

IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF DELAWARE

CODON DEVICES, INC., DUKE)	
UNIVERSITY, and the MASSACHUSETTS)	
INSTITUTE OF TECHNOLOGY,)	
)	C.A. No. _____
Plaintiffs,)	
)	
v.)	DEMAND FOR JURY TRIAL
)	
BLUE HERON BIOTECHNOLOGY, INC.,)	
)	
Defendant.)	

COMPLAINT

Plaintiffs Codon Devices, Inc. (“Codon”), Duke University (“Duke”), and the Massachusetts Institute of Technology (“MIT”) allege as follows:

NATURE OF THE ACTION

1. This is an action arising under the patent laws of the United States (35 U.S.C. § 271 et seq.) based upon Defendant Blue Heron Biotechnology, Inc.’s infringement of four patents owned by Duke and one patent owned by MIT relating generally to the preparation and manufacture of nucleic acids and the assembly of genes from nucleic acids. Exclusive rights in and to each of these five patents have been granted to Codon. Duke, MIT, and Codon seek damages for Defendant’s infringement and a permanent injunction restraining Defendant from further infringement.

PARTIES

2. Plaintiff Codon Devices, Inc. is a Delaware corporation which maintains its principal place of business at One Kendall Square, Building 300, Cambridge, Massachusetts.

3. Plaintiff Duke University is an educational and research institution located in Durham, North Carolina.

4. Plaintiff Massachusetts Institute of Technology is an educational and research institution located in Cambridge, Massachusetts.

5. Upon information and belief, Defendant Blue Heron Biotechnologies, Inc. (“Blue Heron”) is a Delaware corporation which maintains its principal place of business at 22310 20th Avenue SE #100, Bothell, Washington.

JURISDICTION AND VENUE

6. This Court has subject matter jurisdiction under 28 U.S.C. §§ 1331 and 1338.

7. This Court has personal jurisdiction over Defendant because, among other reasons, Defendant resides in this district.

8. Venue is proper in this judicial district under 28 U.S.C. §§ 1391(b) and (c), and § 1400(b).

BACKGROUND

9. Duke is the owner of U.S. Patent No. 5,459,039 (“the ‘039 Patent”), entitled “Methods for mapping genetic mutations.” The ‘039 patent was duly and legally issued to Paul L. Modrich, Shin-San Su, Karin G. Au, and Robert S. Lahue on October 17, 1995, and was assigned to Duke. A true and correct copy of the ‘039 Patent is attached to this Complaint as Exhibit A.

10. Duke is the owner of U.S. Patent No. 5,556,750 (“the ‘750 Patent”), entitled “Methods and kits for fractionating a population of DNA molecules based on the presence or absence of a base-pair mismatch utilizing mismatch repair systems.” The ‘750 patent was duly and legally issued to Paul L. Modrich, Shin-San Su, Karin G. Au, Robert S.

Lahue, Deani L. Cooper, and Leroy Worth, Jr. on September 17, 1996, and was assigned to Duke. A true and correct copy of the '750 Patent is attached to this Complaint as Exhibit B.

11. Duke is the owner of U.S. Patent No. 5,679,522 ("the '522 Patent"), entitled "Methods of analysis and manipulation of DNA utilizing mismatch repair systems." The '522 patent was duly and legally issued to Paul L. Modrich, Shin-San Su, Karin G. Au, Robert S. Lahue, Deani L. Cooper, and Leroy Worth, Jr. on October 21, 1997, and was assigned to Duke. A true and correct copy of the '522 Patent is attached to this Complaint as Exhibit C.

12. Duke is the owner of U.S. Patent No. 5,702,894 ("the '894 Patent"), entitled "Methods of analysis and manipulating of DNA utilizing mismatch repair systems." The '894 patent was duly and legally issued to Paul L. Modrich, Shin-San Su, Karin G. Au, Robert S. Lahue, Deani L. Cooper, and Leroy Worth, Jr. on December 30, 1997, and was assigned to Duke. A true and correct copy of the '894 Patent is attached to this Complaint as Exhibit D.

13. MIT is the owner of U.S. Patent No. 5,750,335 ("the '335 Patent"), entitled "Screening for genetic variation." The '335 patent was duly and legally issued to David K. Gifford on May 12, 1998, and was assigned to MIT. A true and correct copy of the '335 Patent is attached to this Complaint as Exhibit E.

14. Upon information and belief, Defendant manufactures and uses a gene synthesis platform called GeneMaker and provides related services. Defendant's GeneMaker platform, and the methods it performs, are used to synthesize oligonucleotides, hybridize oligonucleotides into duplexes, and assemble genes from the duplexes. Defendant's GeneMaker platform, and the methods it performs, do so using purification techniques, assembly techniques, and other techniques designed to purify, separate and/or detect mismatched nucleic acid duplexes and non-mismatched duplexes for use in the preparation and manufacture of nucleic acids.

15. Exclusive rights in and to the '039, '750, '522, and '894 patents have been licensed to Codon by Duke. Codon is the worldwide exclusive licensee of these patents in a field involving purification, separation and/or detection of mismatched nucleic acid duplexes and non mismatched duplexes for use in the preparation and manufacture of nucleic acids. Pursuant to the license agreement between Codon and Duke, Codon has the right to enforce the '039, '750, '522, and '894 patents within this field. Defendant's GeneMaker platform, the methods it performs, and/or the products of those methods, infringe Codon's exclusive rights in and to the '039, '750, '522, and '894 patents.

16. Exclusive rights in and to the '335 patent have been licensed to Codon by MIT. Codon is the worldwide exclusive licensee of this patent in all fields. Pursuant to the license agreement between Codon and MIT, Codon has the right to enforce the '335 patent. Defendant's GeneMaker platform, the methods it performs, and/or the products of those methods, infringe Codon's exclusive rights in and to the '335 patent.

FIRST CLAIM
(Patent Infringement Of The '039 Patent)

17. On information and belief, Defendant has been and is infringing one or more claims of the '039 Patent, directly and/or indirectly, pursuant to 35 U.S.C. § 271, in connection with its GeneMaker platform.

18. On information and belief, Defendant's infringement of the '039 Patent has been and is willful, and will continue unless enjoined by this Court. Duke and Codon have suffered, and will continue to suffer, irreparable injury as a result of this willful infringement. Pursuant to 35 U.S.C. § 284, Duke and Codon are entitled to damages for infringement and treble damages. Pursuant to 35 U.S.C. § 283, Duke and Codon are entitled to a permanent injunction against further infringement.

19. This case is exceptional and, therefore, Duke and Codon are entitled to attorneys' fees pursuant to 35 U.S.C. § 285.

SECOND CLAIM
(Patent Infringement Of The '750 Patent)

20. On information and belief, Defendant has been and is infringing one or more claims of the '750 Patent, directly and/or indirectly, pursuant to 35 U.S.C. § 271, in connection with its GeneMaker platform.

21. On information and belief, Defendant's infringement of the '750 Patent has been and is willful, and will continue unless enjoined by this Court. Duke and Codon have suffered, and will continue to suffer, irreparable injury as a result of this willful infringement. Pursuant to 35 U.S.C. § 284, Duke and Codon are entitled to damages for infringement and treble damages. Pursuant to 35 U.S.C. § 283, Duke and Codon are entitled to a permanent injunction against further infringement.

22. This case is exceptional and, therefore, Duke and Codon are entitled to attorneys' fees pursuant to 35 U.S.C. § 285.

THIRD CLAIM
(Patent Infringement Of The '522 Patent)

23. On information and belief, Defendant has been and is infringing one or more claims of the '522 Patent, directly and/or indirectly, pursuant to 35 U.S.C. § 271, in connection with its GeneMaker platform.

24. On information and belief, Defendant's infringement of the '522 Patent has been and is willful, and will continue unless enjoined by this Court. Duke and Codon have suffered, and will continue to suffer, irreparable injury as a result of this willful infringement. Pursuant to 35 U.S.C. § 284, Duke and Codon are entitled to damages for infringement and

treble damages. Pursuant to 35 U.S.C. § 283, Duke and Codon are entitled to a permanent injunction against further infringement.

25. This case is exceptional and, therefore, Duke and Codon are entitled to attorneys' fees pursuant to 35 U.S.C. § 285.

FOURTH CLAIM
(Patent Infringement Of The '894 Patent)

26. On information and belief, Defendant has been and is infringing one or more claims of the '894 Patent, directly and/or indirectly, pursuant to 35 U.S.C. § 271, in connection with its GeneMaker platform.

27. On information and belief, Defendant's infringement of the '894 Patent has been and is willful, and will continue unless enjoined by this Court. Duke and Codon have suffered, and will continue to suffer, irreparable injury as a result of this willful infringement. Pursuant to 35 U.S.C. § 284, Duke and Codon are entitled to damages for infringement and treble damages. Pursuant to 35 U.S.C. § 283, Duke and Codon are entitled to a permanent injunction against further infringement.

28. This case is exceptional and, therefore, Duke and Codon are entitled to attorneys' fees pursuant to 35 U.S.C. § 285.

FIFTH CLAIM
(Patent Infringement Of The '335 Patent)

29. On information and belief, Defendant has been and is infringing one or more claims of the '335 Patent, directly and/or indirectly, pursuant to 35 U.S.C. § 271, in connection with its GeneMaker platform.

30. On information and belief, Defendant's infringement of the '335 Patent has been and is willful, and will continue unless enjoined by this Court. MIT and Codon have

suffered, and will continue to suffer, irreparable injury as a result of this willful infringement. Pursuant to 35 U.S.C. § 284, MIT and Codon are entitled to damages for infringement and treble damages. Pursuant to 35 U.S.C. § 283, MIT and Codon are entitled to a permanent injunction against further infringement.

31. This case is exceptional and, therefore, MIT and Codon are entitled to attorneys' fees pursuant to 35 U.S.C. § 285.

WHEREFORE, Duke, MIT, and Codon pray for relief as follows:

PRAYER FOR RELIEF

A. That Defendant be adjudged to have infringed the '039 Patent, '750 Patent, '522 Patent, '894 Patent and '335 Patent;

B. That Defendant, and its officers, agents, servants, employees, attorneys, and those persons in active concert or participation with any of them, be preliminarily and permanently restrained and enjoined from directly or indirectly infringing the '039 Patent, '750 Patent, '522 Patent, '894 Patent and '335 Patent;

C. An accounting for damages by virtue of Defendant's infringement of the '039 Patent, '750 Patent, '522 Patent, '894 Patent and '335 Patent;

D. An award of damages to compensate Duke and Codon for Defendant's infringement of the '039 Patent, '750 Patent, '522 Patent, and '894 Patent, pursuant to 35 U.S.C. § 284, said damages to be trebled because of Defendant's willful infringement;

E. An award of damages to compensate MIT and Codon for Defendant's infringement of the '335 Patent, pursuant to 35 U.S.C. § 284, said damages to be trebled because of Defendant's willful infringement;

F. An assessment of pre-judgment and post-judgment interest and costs against Defendant, together with an award of such interest and costs, in accordance with 35 U.S.C. § 284;

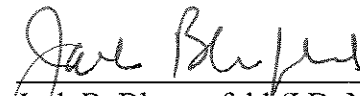
G. That Defendant be directed to pay Duke, MIT, and Codon's attorneys' fees incurred in connection with this lawsuit pursuant to 35 U.S.C. § 285; and

H. That Duke, MIT, and Codon have such other and further relief as this Court may deem just and proper.

JURY DEMAND

Duke, MIT, and Codon demand a trial by jury, pursuant to Fed. R. Civ. P. 38(b), on all disputed issues.

MORRIS, NICHOLS, ARSHT & TUNNELL LLP



(Jack B. Blumenfeld (I.D. No. 1014)
Leslie A. Polizoti (I.D. No. 4299)
1201 North Market Street
P.O. Box 1347
Wilmington, DE 19899-1347
(302) 658-9200
jblumenfeld@mnat.com

*Attorneys for Plaintiffs
Codon Devices, Inc., Duke University
and Massachusetts Institute of Technology*

Of Counsel:

Edward R. Reines
Nicholas A. Brown
Rip J. Finst
WEIL, GOTSHAL & MANGES LLP
201 Redwood Shores Parkway
Redwood Shores, CA 94065
(650) 802-3000

March 14, 2007

EXHIBIT A



US005459039A

United States Patent

[19]

[11] **Patent Number:** **5,459,039****Modrich et al.**[45] **Date of Patent:** **Oct. 17, 1995**[54] **METHODS FOR MAPPING GENETIC MUTATIONS**[75] Inventors: **Paul Modrich**, Durham, N.C.;
Shin-San Su, Cambridge, Mass.; **Karin G. Au**; **Robert S. Lahue**, both of Durham, N.C.[73] Assignee: **Duke University**, Durham, N.C.[21] Appl. No.: **334,612**[22] Filed: **Mar. 18, 1994****Related U.S. Application Data**

[63] Continuation of Ser. No. 2,529, Jan. 11, 1993, abandoned, which is a continuation of Ser. No. 350,983, May 12, 1989, abandoned.

[51] Int. Cl.⁶ **C12Q 1/68**[52] U.S. Cl. **435/6; 935/77; 935/78; 436/501**[58] Field of Search **435/6; 436/501; 935/77, 78**[56] **References Cited****U.S. PATENT DOCUMENTS**

4,794,075 12/1988 Ford et al. 435/6

FOREIGN PATENT DOCUMENTS

9322457 11/1993 WIPO .

OTHER PUBLICATIONS

Lu et al. Cold Spring Harbor Symp. Quant. Biol. 49, 589-96 (1984).

Lu et al. Cell, vol. 54, 805-812, Sep. 9, 1988.

Adams et al. The Biochemistry of Nucleic Acids Chapman & Hall, 1986, pp. 221-223.

Hennighausen et al. Guide to Molec. Cloning Tech, Berger et al. editors Academic Press, Inc. 1987 pp. 721-735.

Quinones et al. Mol Gen Genet. (1988) 211: 106-112.

Su et al. Proc Natl Acad Sci USA vol 83 pp. 5057-5061 Jul. 1986.

Cotton et al. Proc Natl Acad Sci. USA vol. 85 pp. 4397-4401 Jun. 1988.

Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1986, NIH Grant, Abstract (Source: CRISP).
 Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1987, NIH Grant, Abstract (Source: CRISP).
 Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1988, NIH Grant, Abstract (Source: CRISP).
 Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1989, NIH Grant, Abstract (Source: CRISP).
 Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1990, NIH Grant, Abstract (Source: CRISP).
 Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1991, NIH Grant, Abstract (Source: CRISP).
 Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1992, NIH Grant, Abstract (Source: CRISP).
 Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1993, NIH Grant, Abstract (Source: CRISP).
 Modrich, "Enzymology of Eukaryotic DNA Mismatch Repair", 1991, NIH Grant, Abstract (Source: CRISP).
 Modrich, "Enzymology of Eukaryotic DNA Mismatch Repair", 1992, NIH Grant, Abstract (Source: CRISP).
 Modrich, "Enzymology of Eukaryotic DNA Mismatch Repair", 1993, NIH Grant, Abstract (Source: CRISP).
 Su et al, Mismatch Specificity of Methyl-directed DNA Mismatch in Vitro, The Journal of Biological Chemistry, 1988, pp. 6829-6835.

Primary Examiner—Margaret Parr
Assistant Examiner—Carla Myers
Attorney, Agent, or Firm—Lyon & Lyon

[57]

ABSTRACT

The present invention relates to a method for detecting base sequence differences within homologous regions of two DNA molecules comprising the steps of contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting DNA duplexes with a protein that recognizes substantially all base pair mismatches under conditions such that the protein forms specific complexes with its cognate mispairs, and detecting the resulting DNA:protein complexes by a suitable analytical method. Also disclosed are protein components of DNA mismatch correction systems and the use of these components in methods for genetic mapping.

6 Claims, 3 Drawing Sheets

U.S. Patent

Oct. 17, 1995

Sheet 1 of 3

5,459,039

V 5'-AAGCTTTCGAG Hind III
C 3'-TTCGAGAGCTC Xho I

FIG. 1.

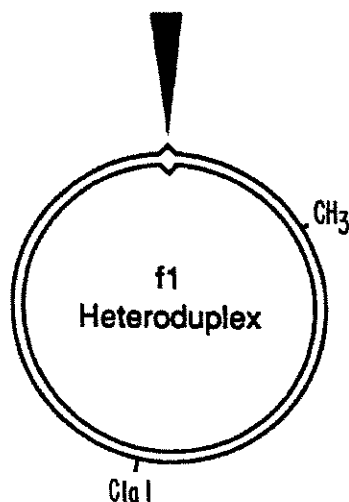
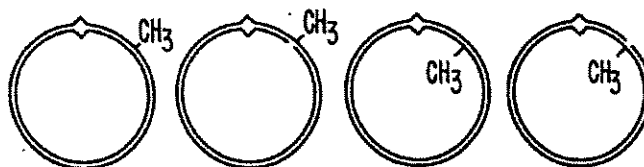
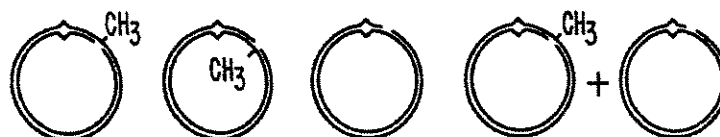


FIG. 4.



REACTION CONDITIONS	REPAIR (fmol/20 min)			
COMPLETE	15 (<1)	17 (<1)	8 (<1)	10 (<1)
- Mut H	<1	18	1	9
- Mut L	<1	<1	<1	<1
- Mut S	<1	<1	<1	1
- SSB	2	<1	<1	<1
- pol III holoenzyme	<1	<1	<1	<1

FIG. 5.



LIGASE	MutH	REPAIR (fmol/20 min)				
-	-	19 (<1)	9 (<1)	11 (<1)	19 (<1)	9 (<1)
+	-	2	<1	1	2	1
+	+	20	7	2	15	1

U.S. Patent

Oct. 17, 1995

Sheet 2 of 3

5,459,039

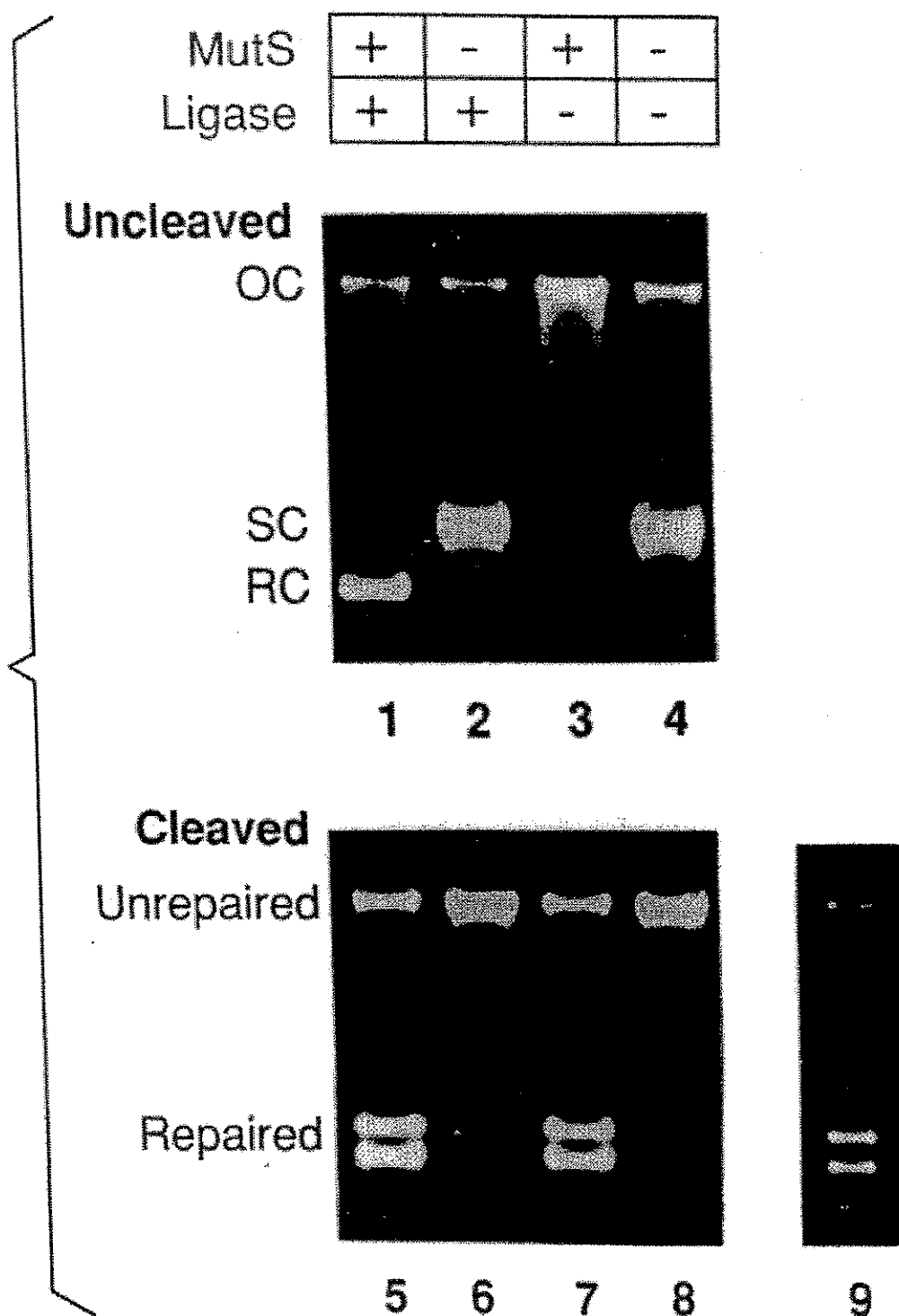


FIG. 2.

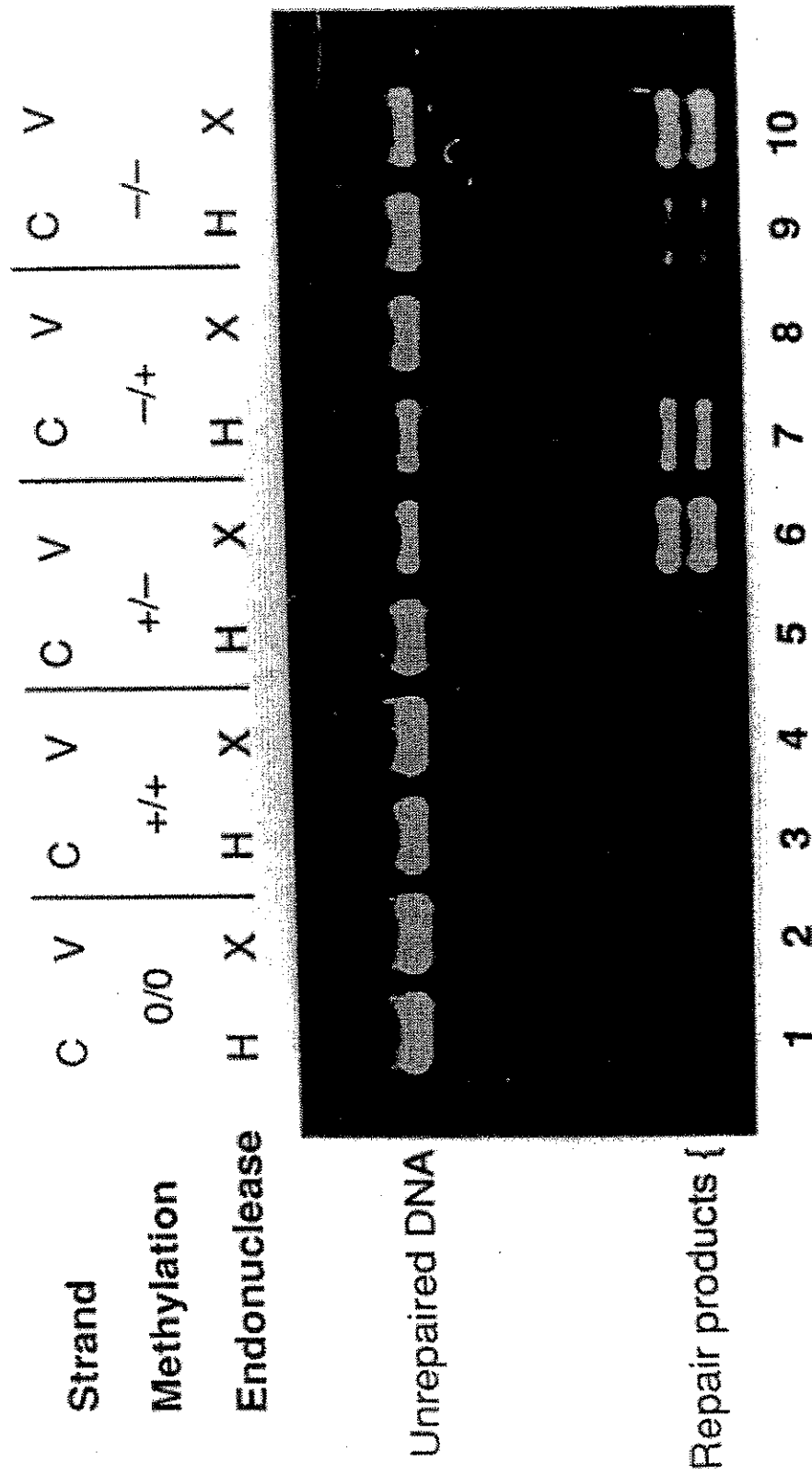
U.S. Patent

Oct. 17, 1995

Sheet 3 of 3

5,459,039

FIG. 3.



5,459,039

1

METHODS FOR MAPPING GENETIC MUTATIONS

This invention was made with Government support under Contract No. GM-23719 awarded by the National Institutes of Health. The Government has certain rights in the invention. This application is a continuation of application Ser. No. 08/002,529, filed Jan. 11, 1993, now abandoned, which is a continuation application Ser. No. 07/350,983, filed May 12, 1989, abandoned.

FIELD OF THE INVENTION

The present invention relates to methods for mapping genetic differences among deoxyribonucleic acid ("DNA") molecules, especially mutations involving a difference in a single base between the base sequences of two homologous DNA molecules. More specifically, this invention relates to such mapping methods which employ proteins that recognize and correct mismatched DNA base pairs in double-stranded DNA. This invention also relates to the manufacture and use of certain novel products enabled by the identification and isolation of proteins that are components of mismatched base pair recognition and correction systems.

BACKGROUND OF THE INVENTION

Mapping of genetic differences between individuals is of growing importance for both forensic and medical applications. For example, DNA "fingerprinting" methods are being applied for identification of perpetrators of crimes where even small amounts of blood or sperm are available for analysis. Biological parents can also be identified by comparing DNAs of a child and a suspected parent using such means. Further, a number of inherited pathological conditions may be diagnosed before onset of symptoms, even in utero, using methods for structural analyses of DNA. Finally, it is notable that a major international effort to physically map and, ultimately, to determine the sequence of bases in the DNA encoding the entire human genome is now underway and gaining momentum in both institutional and commercial settings.

DNA molecules are linear polymers of subunits called nucleotides. Each nucleotide comprises a common cyclic sugar molecule, which in DNA is linked by phosphate groups on opposite sides to the sugars of adjoining nucleotides, and one of several cyclic substituents called bases. The four bases commonly found in DNAs from natural sources are adenine, guanine, cytosine and thymine, hereinafter referred to as A, G, C and T, respectively. The linear sequence of these bases in the DNA of an individual encodes the genetic information that determines the heritable characteristics of that individual.

In double-stranded DNA, such as occurs in the chromosomes of all cellular organisms, the two DNA strands are entwined in a precise helical configuration with the bases projecting inward and so aligned as to allow interactions between bases from opposing strands. The two strands are held together in precise alignment mainly by hydrogen bonds which are permitted between bases by a complementarity of structures of specific pairs of bases. This structural complementarity is determined by the chemical natures and locations of substituents on each of the bases. Thus, in double-stranded DNA, normally each A on one strand pairs with a T from the opposing strand, and, likewise, each G with an opposing C.

When a cell undergoes reproduction, its DNA molecules

2

are replicated and precise copies are passed on to its descendants. The linear base sequence of a DNA molecule is maintained in the progeny during replication in the first instance by the complementary base pairings which allow each strand of the DNA duplex to serve as a template to align free nucleotides with its polymerized nucleotides. The complementary nucleotides so aligned are biochemically polymerized into a new DNA strand with a base sequence that is entirely complementary to that of the template strand.

Occasionally, an incorrect base pairing does occur during replication, which, after further replication of the new strand, results in a double-stranded DNA offspring with a sequence containing a heritable single base difference from that of the parent DNA molecule. Such heritable changes are called genetic mutations, or more particularly in the present case, "single base pair" or "point" mutations. The consequences of a point mutation may range from negligible to lethal, depending on the location and effect of the sequence change in relation to the genetic information encoded by the DNA.

The bases A and G are of a class of compounds called purines, while T and C are pyrimidines. Whereas the normal base pairings in DNA (A with T, G with C) involve one purine and one pyrimidine, the most common single base mutations involve substitution of one purine or pyrimidine for the other (e.g., A for G or C for T), a type of mutation referred to as a "transition". Mutations in which a purine is substituted for a pyrimidine, or vice versa, are less frequently occurring and are called "transversions". Still less common are point mutations comprising the addition or loss of a single base arising in one strand of a DNA duplex at some stage of the replication process. Such mutations are called single base "insertions" or "deletions", respectively, and are also known as "frameshift" mutations, due to their effects on translation of the genetic code into proteins. Larger mutations affecting multiple base pairs also do occur and can be important in medical genetics, but their occurrences are relatively rare compared to point mutations.

Mapping of genetic mutations involves both the detection of sequence differences between DNA molecules comprising substantially identical (i.e., homologous) base sequences, and also the physical localization of those differences within some subset of the sequences in the molecules being compared. In principle, it is possible to both detect and localize limited genetic differences, including point mutations within genetic sequences of two individuals, by directly comparing the sequences of the bases in their DNA molecules. In practice, however, direct DNA sequencing has highly restricted usefulness for mapping mutations due to the major time and effort required to determine the sequence of even one DNA fragment comprising a few hundred base pairs. Typically, a single functional unit of genetic information, a gene, may be encoded in tens of thousands of base pairs of human chromosomal DNA. Thus comparing the sequence of a complete gene from one individual with that of another by direct DNA sequencing involves analyses of multiple short fragments of that gene, requiring many months if not years of effort. It may also be noted that there are estimated to be hundreds of thousands of genes in the entire human gene complement or genome, as it is called, any one of which may be involved in some genetically determined disease.

Accordingly, several simpler methods for detecting differences between DNA sequences have been developed which although providing less direct information about base sequence differences, nevertheless do yield useful observations under limited circumstances. For example, some pairs

5,459,039

3

of single-stranded DNA fragments with sequences differing in a single base may be distinguished by their different migration rates in electric fields, as in denaturing gradient gel electrophoresis. This method does not detect all the possible single-base differences between DNA fragments and is restricted to fragments comprising at most a few hundred base pairs. Further, it is technically difficult to generate consistent analyses using this method. Thus this approach has extremely limited utility for detection and localization of single base sequence differences between DNAs encoding whole genes.

DNA restriction systems found in bacteria, for example, comprise proteins which generally recognize specific sequences in double-stranded DNA composed of 4 to 6 or more base pairs. In the absence of certain modifications (e.g., a covalently attached methyl group) at definite positions within the restriction recognition sequence, endonuclease components of the restriction system will cleave both strands of a DNA molecule at specific sites within or near the recognition sequence. Such short recognition sequences occur by chance in all natural DNA sequences, once in every few hundred or thousand base pairs, depending on the recognition sequence length. Thus, digestion of a DNA molecule with various restriction endonucleases, followed by analyses of the sizes of the resulting fragments (e.g., by gel electrophoresis), may be used to generate a physical map ("fingerprint") of the locations in a DNA molecule of selected short sequences.

It is well known in the art that comparisons of such restriction maps of two homologous DNA sequences can reveal differences within those specific sequences that are recognized by those restriction enzymes used in the available maps. Restriction map comparisons may localize any detectable differences within limits defined ultimately by the resolving power of DNA fragment size determination, essentially within about the length of the restriction recognition sequence under certain conditions of gel electrophoresis. To achieve such resolution in location of a point mutation by restriction mapping, however, all fragments resulting from digestion with each restriction nuclease must be within a range of distinguishable sizes, usually below an upper limit of between 10 and 20 thousand base pairs (kbp), and preferably less than one kbp, using standard gel electrophoresis techniques. Since each different restriction enzyme scans only a fraction of a percent of all the sequences in any DNA molecule, literally thousands of analyses with thousands of different enzymes would be needed to completely compare two DNAs encoding even one gene, assuming that enzymes recognizing all possible 4 to 6 base sequences were known, which they are not.

In practice, selected heritable differences in restriction fragment lengths (i.e., restriction fragment length polymorphisms, "RFLP's") have been extremely useful, for instance, for generating physical maps of the human genome on which genetic defects may be located with a relatively low precision of hundreds or, sometimes, tens of thousands of base pairs. Typically, RFLPs are detected in human DNA isolated from small tissue or blood samples by using radioactively labeled DNA fragments complementary to the genes of interest. These "probes" are allowed to form DNA duplexes with restriction fragments of the human DNA after separation by electrophoresis, and the resulting radioactive duplex fragments are visualized by exposure to photographic (e.g., X-ray sensitive) film, thereby allowing selective detection of only the relevant gene sequences amid the myriad of others in the genomic DNA.

When the search for DNA sequence differences can be

4

confined to specific regions of known sequence, the recently developed "polymerase chain reaction" ("PCR") technology can be used to reduce the amount of effort needed to detect and locate a single base difference as compared to the usual DNA sequencing approach which requires molecular cloning of the DNA fragment of interest. Briefly, this method utilizes short DNA fragments complementary to sequences on either side of the location to be analyzed to serve as points of initiation for DNA synthesis (i.e., "primers") by purified DNA polymerase. The resulting cyclic process of DNA synthesis results in massive biochemical amplification of the sequences selected for analysis, which then may be easily detected and, if desired, further analyzed, for example, by restriction mapping or direct DNA sequencing methods. In this way, selected regions of a human gene comprising a few kbp may be amplified and examined for sequence variations, but only in cases where sequences spanning a particular location of interest are known.

In clinical practice, the PCR method is of limited utility, for example, in detection of known heritable variants of selected human genes which differ by only one or a few specific base pairs (i.e., allelic forms a gene). For example, the human β -globin gene comprises several alleles that can be distinguished by this approach; but the overall utility is highly limited, particularly when faced with a need to detect sequence differences which may be scattered over large stretches of a gene, as in the diagnosis of conditions resulting from frequent new mutational events in human populations, in the Lesch-Nyan syndrome, for example.

Another known method for detecting and localizing single base differences within homologous DNA molecules involves the use of a radiolabeled RNA fragment with base sequence complementary to one of the DNAs and a nuclease that recognizes and cleaves single-stranded RNA. The structure of RNA is highly similar to DNA, except for a different sugar and the presence of uracil (U) in place of T; hence, RNA and DNA strands with complementary sequences can form helical duplexes ("DNA:RNA hybrids") similar to double-stranded DNA, with base pairing between A's and U's instead of A's and T's. It is known that the enzyme ribonuclease A ("RNase A") can recognize some single pairs of mismatched bases (i.e., "base mispairs") in DNA:RNA hybrids and can cleave the RNA strand at the mispair site. Analysis of the sizes of the products resulting from RNase A digestion allows localization of single base mismatches, potentially to the precise sequence position, within lengths of homologous sequences determined by the limits of resolution of the RNA sizing analysis (Myers, R. M. et al., 1985, *Science*, 230, 1242-1246). RNA sizing is performed in this method by standard gel electrophoresis procedures used in DNA sequencing, an approach which limits the practical resolution to mapping of single base mispairs in a DNA:RNA hybrid comprising an RNA of only several hundred nucleotides. Moreover, this RNase A method requires preparing complementary RNA probes from each DNA sequence to be examined, which requires more work and is more technically demanding than methods using only DNA (such as restriction mapping). Further, RNase A does not efficiently recognize all possible mispairings of DNA and RNA bases, resulting in a significant inefficiency in detection of all point differences between DNA sequences.

It has also been reported that S1 nuclease, an endonuclease specific for single-stranded nucleic acids, can recognize and cleave limited regions of mismatched base pairs in DNA:DNA or DNA:RNA duplexes. Therefore, it has been suggested that S1 nuclease could be used to map single base pair differences between DNA molecules by sizing of cleav-

5,459,039

5

age fragments. However, more extensive analysis of this enzyme has established that a mismatch of at least about 4 consecutive base pairs actually is generally required for recognition and cleavage of a duplex by S1 nuclease, thus precluding its use for detection of any point mutations.

Thus, none of the available methods for comparing the base sequences of DNAs, other than direct sequencing, can efficiently detect and localize all possible single base differences. Further, all of these methods, including especially DNA sequencing, require substantial labor and repetitive analyses with various sequence specific reagents (e.g., multiple nucleases or short nucleic acid strands) to detect all single base differences within two specimens of a single human gene.

Hence, there is a need for simpler and more efficient approaches, both for detecting and for localizing genetic differences between DNA sequences to facilitate both clinical diagnoses and forensic investigations. In particular, the observations above indicate a specific need for simpler and more efficient methods and reagents for detection of any possible single base differences between long DNA sequences, for example, between a complete gene from one individual and the entire genome of another. There is also a further need for simpler methods for localization of any possible single base differences within the sequences of homologous regions of long DNA molecules such as those encoding one or more complete genes and comprising several kbp of DNA.

The present invention contemplates the use of certain proteins that recognize mismatched base pairs in double-stranded DNA (and, therefore, are called "mismatch recognition proteins") in defined systems for detecting and mapping point mutations in DNAs. Accordingly, it is an object of the present invention to provide methods for using such mismatch recognition proteins, alone or in combination with other proteins, for detecting and localizing single base differences between DNA molecules, particularly those DNAs comprising several kbp. Additionally, it is an object of this invention to develop modified forms of mismatch recognition proteins to further simplify methods for identifying specific bases which differ between DNAs.

Enzymatic systems capable of recognition and correction of base pairing errors within the DNA helix have been demonstrated in bacteria, fungi and mammalian cells, but the mechanisms and functions of mismatch correction are best understood in *Escherichia coli*. Of the several mismatch repair systems that have been identified in *E. coli*, the most relevant here is the methyl-directed pathway for repair of DNA biosynthetic errors. The fidelity of DNA replication in *E. coli* is enhanced 100-1000 fold by this postreplication mismatch correction system. This system processes base pairing errors within the helix in a strand-specific manner by exploiting patterns of DNA methylation. Since DNA methylation is a postsynthetic modification, newly synthesized strands temporarily exist in an unmethylated state, with the transient absence of adenine methylation on GATC sequences directing mismatch correction to new DNA strands within the hemimethylated duplexes.

In vivo analyses in *E. coli* have shown that selected examples of each of the different mismatches are subject to correction with different efficiencies. G-T, A-C, G-G and A-A mismatches are typically subject to efficient repair. A-G, C-T, T-T and C-C are weaker substrates, but well repaired exceptions exist within this class. It is thought that the sequence environment of a mismatched base pair may be an important factor in determining the efficiency of repair in

6

vivo. The mismatch correction system is also capable in vivo of correcting differences between duplexed strands involving a single base insertion or deletion. Further, genetic analyses have demonstrated that the mismatch correction process requires intact genes for several proteins, including the products of the *mutH*, *mutL* and *mutS* genes, as well as DNA helicase II and single-stranded DNA binding protein (SSB).

The present inventors have been seeking to identify and isolate specific proteins that are required for correction of mismatched base pairs and to understand the specific biochemical functions of these mismatch correction system components. The products of the *mutH* and *mutS* genes have been purified to near homogeneity in biologically active form. Analysis of the *MutH* protein has suggested that it functions in strand discrimination by incising the unmethylated DNA strand at GATC sites. The isolated *MutS* protein has been shown to recognize four of the eight possible mismatched base pairs (specifically, G-T, A-C, A-G and C-T mispairs; Su, S. -S. and Modrich, P., 1986, *Proc. Nat. Acad. Sci. U.S.A.*, 84, 5057-5061). The hierarchy of apparent affinities of isolated *MutS* protein for the particular examples of the four mispairs tested in these studies did not correlate well with in vivo efficiencies of mismatch correction. Hence, these studies left undetermined whether or not additional proteins, acting alone or in concert with *MutS*, are required for or influence the recognition of other base mispairs.

SUMMARY OF THE INVENTION

It has now been discovered that a single DNA base mismatch recognition protein can form specific complexes with any of the eight possible mismatched base pairs embedded in an otherwise homologous DNA duplex. It has also been revealed that another mismatch recognition protein can recognize only one specific base pair mismatch, A-G, and in so doing, it chemically modifies a nucleotide at the site of the mismatch. In addition, defined in vitro systems have been established for carrying out methyl-directed mismatch repair processes. Accordingly, the present invention contemplates the use of such mismatch recognition proteins and related correction system components to detect and to localize point mutations in DNAs.

For clarity in the following discussion, it will be useful to point out here certain distinctions related to the fact that some proteins that recognize DNA base mispairs are merely DNA binding proteins, while others modify the DNA as a consequence of mismatch recognition. Notwithstanding the fact that in the latter situation the protein modifying the DNA may be associated with the DNA only transiently, hereinafter, whether a mismatch recognition protein is capable of DNA binding only or also of modifying DNA, whenever it is said that a protein recognizes a DNA mismatch, this is equivalent to saying that it "forms specific complexes with" or "binds specifically to" that DNA mismatch in double-stranded DNA. In the absence of express reference to modification of DNA, reference to DNA mismatch recognition does not imply consequent modification of the DNA. Further, the phrase "directs modification of DNA" includes both cases wherein a DNA mismatch recognition protein has an inherent DNA modification function (e.g., a glycosylase) and cases wherein the mismatch recognition protein merely forms specific complexes with mispairs, which complexes are then recognized by other proteins that modify the DNA in the vicinity of the complex. Finally, it should be noted in the following discussion that those DNA base mispairs (e.g.,

5,459,039

7

A-G or C-C) which are recognized by a given protein are referred to as the "cognate" base mispairs for that protein.

Accordingly, the present invention relates to a method for detecting base sequence differences within homologous regions of two DNA molecules comprising the steps of contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting DNA duplexes with a protein that recognizes substantially all base pair mismatches under conditions such that the protein forms specific complexes with its cognate mispairs, and detecting the resulting DNA:protein complexes by a suitable analytical method.

In the practice of a preferred embodiment of this aspect of this invention, the mispair recognition protein is the product of the *mutS* gene of *E. coli* or another functionally homologous protein, and an advantageous analytical method for detecting the DNA:protein complex comprises the steps of contacting the DNA:protein complexes with a selectively adsorbent agent, such as a membranous nitrocellulose filter, under conditions such that protein:DNA complexes are retained on the agent while DNA not complexed with protein is not retained, and measuring the amount of DNA in the retained complexes. Other suitable analytical methods for detecting the DNA:protein complex are disclosed.

In addition to methods designed merely to detect base sequence differences between DNAs, this invention further relates to a method for both detecting and localizing individual base sequence differences within homologous regions of two DNA molecules comprising the steps of contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting double-stranded DNA duplexes with a protein that recognizes at least one base mispair under conditions such that the protein forms specific complexes with its cognate mispairs and thereby directs modification of at least one strand of the DNA in the resulting DNA:protein complexes in the vicinity of the DNA:protein complex, and determination of the location of the resulting DNA modification relative to a known sequence within the homologous regions of the DNAs by a suitable analytical method.

In the practice of one embodiment of this aspect of this invention, the mispair recognition protein is the product of the *mutS* gene of *E. coli* or is another functionally homologous protein; the step in which the DNA is modified in the vicinity of the DNA:protein complex further comprises contacting the DNA:MutS protein complex with a defined *E. coli* DNA mismatch correction system under conditions such that single-stranded gaps are produced in the vicinity of the complexed protein; and the method for determining the locations of these single-stranded gaps within the DNA duplex comprises the steps of cleaving the DNA with a single-strand specific endonuclease and at least one restriction endonuclease, and comparing the electrophoretic mobilities of the resulting modified DNA fragments with DNA restriction fragments not contacted with the defined mismatch correction system. Suitable single-strand specific endonucleases include the S1 single-strand specific nuclease, for example, or other functionally similar nucleases well known in the art.

The present invention further relates, in part, to forms of mispair recognition proteins which have been altered to provide an inherent means for modifying at least one strand of the DNA duplex in the vicinity of the bound mispair recognition protein.

8

In a principal embodiment of this aspect of this invention, the altered mispair recognition protein is the modified product of the *mutS* gene of *E. coli* or is another functionally homologous modified protein to which is attached an hydroxyl radical cleaving function; and the DNA modification step in the DNA mispair localization method further comprises contacting this modified protein with the DNA in under conditions such that the radical cleaving function cleaves at least one strand of the DNA in the vicinity of the protein. Additional altered forms of mispair recognition proteins that modify at least one strand of the DNA in a DNA:protein complex in the vicinity of the bound protein are disclosed.

The present invention also comprises another *E. coli* DNA mispair recognition protein that recognizes only A-G mispairs without any apparent requirement for hemimethylation. This protein, the product of the *mutY* gene, is a glycosylase which specifically removes the adenine from an A-G mispair in a DNA duplex. Accordingly, this *MutY* protein is useful for the specific detection of A-G mispairs according to the practice of the present invention.

BRIEF DESCRIPTION OF THE FIGURES

FIG. 1. Heteroduplex substrate for in vitro mismatch correction. Each substrate used in this study is a 6440-bp, covalently closed, circular heteroduplex that is derived from bacteriophage ϕ 1 and contains a single base-base mismatch located within overlapping recognition sites for two restriction endonucleases at position 5632. In the example shown a G-T mismatch resides within overlapping sequences recognized by Hind III and Xho I endonucleases. Although the presence of the mispair renders this site resistant to cleavage by either endonuclease, repair occurring on the complementary (c) DNA strand yields an A-T base pair and generates a Hind III-sensitive site, while correction on the viral (v) strand results in a G-C pair and Xho I-sensitivity. The heteroduplexes also contain a single d(GATC) sequence 1024 base pairs from the mismatch (shorter path) at position 216. The state of strand methylation at this site can be controlled, thus permitting evaluation of the effect of DNA methylation on the strand specificity of correction.

FIG. 2. Requirement for DNA ligase in mismatch correction. Hemimethylated G-T heteroduplex DNA [FIG. 1, 0.6 μ g, d(GATC) methylation on the complementary DNA strand] was subjected to mismatch repair under reconstituted conditions in a 60 μ l reaction (Table 3, closed circular heteroduplex), or in 20 μ l reactions (0.2 μ g of DNA) lacking *MutS* protein or ligase, or lacking both activities. A portion of each reaction (0.1 μ g of DNA) was treated with EDTA (10 mM final concentration) and subjected to agarose gel electrophoresis in the presence of ethidium bromide (1.5 μ g/ml; top panel, lanes 1-4). Positions are indicated for the unreacted, supercoiled substrate (SC), open circles containing a strand break (OC) and covalently closed, relaxed circular molecules (RC). A second sample of each reaction containing 0.1 μ g of DNA was hydrolyzed with Xho I and Cla I endonucleases (FIG. 1) to score G-T \rightarrow G-C mismatch correction and subjected to electrophoresis in parallel with the samples described above (bottom panel, lanes 5-8). The remainder of the complete reaction (0.4 μ g DNA, corresponding to the sample analyzed in lane 1) was made 10 mM in EDTA, and subjected to electrophoresis as described above. A gel slice containing closed circular, relaxed molecules was excised and the DNA eluted. This sample was cleaved with Xho I and Cla I and the products analyzed by electrophoresis (lane 9).

5,459,039

9

FIG. 3. Methyl-direction of mismatch correction in the purified system. Repair reactions with the G-T heteroduplex (FIG. 1) were performed as described in Table 3 (closed circular heteroduplex) except that reaction volumes were 20 μ l (0.2 μ g of DNA) and the incubation period was 60 minutes. The reactions were heated to 55° for 10 minutes and each was divided into two portions to test strand specificity of repair. G-T→A-T mismatch correction, in which repair occurred on the complementary (c) DNA strand, was scored by cleavage with Hind III and Cla I endonucleases, while hydrolysis with Xho I and Cla I were used to detect G-T→G-C repair occurring on the viral (v) strand. Apart from the samples shown in the left two lanes, all heteroduplexes were identical except for the state of methylation of the single d(GATC) sequence at position 216 (FIG. 1). The state of modification of the two DNA strands at this site is indicated by + and - notation. The G-T heteroduplex used in the experiment shown in the left two lanes (designated 0/0) contains the sequence d(GATT) instead of d(GATC) at position 216, but is otherwise identical in sequence to the other substrates.

FIG. 4. Strand-specific repair of heteroduplexes containing a single strand scission in the absence of MutH protein. Hemimethylated G-T heteroduplex DNAs (FIG. 1, 5 μ g) bearing d(GATC) modification on the viral or complementary strand were subjected to site-specific cleavage with near homogeneous MutH protein. Because the MutH-associated endonuclease is extremely weak in the absence of other mismatch repair proteins, cleavage at d(GATC) sites by the purified protein requires a MutH concentration 80 times that used in reconstitution reactions. After removal of MutH by phenol extraction, DNA was ethanol precipitated, collected by centrifugation, dried under vacuum, and resuspended in 10 mM Tris-HCl (pH 7.6), 1 mM EDTA. Mismatch correction of MutH-incised and covalently closed, control heteroduplexes was performed as described in the legend to Table 2 except that ligase and NAD⁺ were omitted. Outside and inside strands of the heteroduplexes depicted here correspond to complementary and viral strands respectively. Values in parentheses indicate repair occurring on the methylated, continuous DNA strand. The absence of MutH protein in preparations of incised heteroduplexes was confirmed in two ways. Preparations of incised molecules were subject to closure by DNA ligase (>80%) demonstrating that MutH protein does not remain tightly bound to incised d(GATC) sites. Further, control experiments in which each MutH incised heteroduplex was mixed with a closed circular substrate showed that only the open circular form was repaired if MutH protein was omitted from the reaction whereas both substrates were corrected if MutH protein was present (data not shown).

FIG. 5. Requirements for MutH protein and a d(GATC) sequence for correction in the presence of DNA ligase. Hemimethylated G-T heteroduplexes incised on the unmethylated strand at the d(GATC) sequence were prepared as described in the legend to FIG. 4. A G-T heteroduplex devoid of d(GATC) sites (FIG. 4) and containing a single-strand break within the complementary DNA strand at the Hinc II site (position 1) was constructed as described previously. Mismatch correction assays were performed as described in Table 3, with ligase (20 ng in the presence of 25 μ M NAD⁺) and MutH protein (0.26 ng) present as indicated. Table entries correspond to correction occurring on the incised DNA strand, with parenthetical values indicating the extent of repair on the continuous strand. Although not shown, repair of the nicked molecule lacking a d(GATC) sequence (first entry of column 3) was reduced more than an

10

order of magnitude upon omission of MutL, MutS, SSB or DNA polymerase III holoenzyme.

DESCRIPTION OF SPECIFIC EMBODIMENTS

The present invention relates to a method for detecting base sequence differences within homologous regions of two DNA molecules comprising the steps of contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting DNA duplexes with a protein that recognizes at least one base pair mismatch under conditions such that the protein forms specific complexes with its cognate mispairs, and detecting the resulting DNA:protein complexes by a suitable analytical method.

In the practice of this method, the two DNA molecules to be compared may comprise natural or synthetic sequences encoding up to the entire genome of an organism, including man, which can be prepared by well known procedures. Detection of bases sequence differences according to this method of this invention does not require cleavage (by a restriction nuclease, for example) of either of the two DNAs, although it is well known in the art that rate of base pair formation between complementary single-stranded DNA fragments is inversely related to their size. This detection method requires that base sequence differences to be detected lie within a region of homology constituting at least about 14 consecutive base pairs of homology between the two DNA molecules, which is about the minimum number of base pairs generally required to form a stable DNA duplex. Either one or both of the strands of the first DNA may be selected for examination, while at least one strand of the second DNA complementary to a selected first DNA strand must be used. The DNA strands, particularly those of the second DNA, advantageously may be radioactively labeled to facilitate direct detection, according to procedures well known in the art.

Methods and conditions for contacting the DNA strands of the two DNAs under conditions such that base pairing occurs are also widely known in the art.

In the practice of a principal embodiment of this aspect of this invention, the mispair recognition protein is the product of the mutS gene of *E. coli*. Preparation of this protein substantially free of other proteins has been reported previously (Su, S.-S. and Modrich, P., 1986, *Proc. Nat. Acad. Sci. U.S.A.*, 84, 5057-5061, which is hereby incorporated herein by reference).

The surprising ability of the MutS protein to recognize examples of all eight single base pair mismatches within double-stranded DNA, even including C-C mispairs which do not appear to be corrected *in vivo*, is demonstrated by the fact that MutS protein protects DNA regions containing each mismatch from hydrolysis by DNase I (i.e., by "DNase I footprint" analyses), as recently reported (Su, S.-S., et al., 1988, *J. Biol. Chem.*, 263, 6829-6835). The affinity of MutS protein for the different mispairs that have been tested varies considerably. Local sequence environment may also affect the affinity of the MutS protein for any given base mispair; in other words, for example, the affinity for two specific cases of A-C mispairs, which are surrounded by different sequences, may not be the same. Nevertheless, no examples of base mispairs have been found that are not recognized by isolated MutS protein. Accordingly, it is believed that this method of this invention detects substantially all possible single base differences between homologous regions of any two DNA molecules.

5,459,039

11

It should be particularly noted that the DNA duplexes which MutS recognizes are not required to contain GATC sequences and, hence, they do not require hemimethylation of A's in GATC sequences, the specific signal for the full process of mispair correction in vivo; therefore, use of MutS in this method allows recognition of a DNA base mispair in DNAs lacking such methylation, for instance, DNAs isolated from human tissues.

A protein which appears to be functionally and in part, at least, structurally homologous to the *E. coli* MutS protein has also been discovered in a methyl-directed mispair correction system in *Salmonella typhimurium* bacteria (Pang et al., 1985, *J. Bacteriol.*, 163, 1007-1015). The gene for this protein has been shown to complement *E. coli* strains with mutations inactivating the *mutS* gene and the amino acid sequence of its product shows homology with that of the *E. coli* MutS protein. Accordingly, this *S. typhimurium* protein is also believed to be suitable for the practice of this aspect of the present invention. Other organisms, including man, are known to possess various systems for recognition and repair of DNA mispairs, which, as one skilled in the art would appreciate, comprise mispair recognition proteins functionally homologous to the MutS protein. Accordingly, it is believed that such DNA base mispair recognition proteins are also suitable for use in the present invention.

In the practice of a preferred embodiment of this aspect of this invention, an advantageous analytical method for detecting the DNA:protein complex comprises the steps of contacting the DNA:protein complexes with a selectively adsorbent agent, such as a membranous nitrocellulose filter, under conditions such that protein:DNA complexes are retained on the agent while DNA not complexed with protein is not retained, and measuring the amount of DNA in the retained complexes. Absent radioactive labeling of at least one strand used to form the DNA duplexes, the DNA in complexes on the filter may be detected by any of the usual means in the art for detection of DNA on a solid substrate, including annealing with complementary strands of radioactive DNA.

The nitrocellulose filter method for detecting complexes of MutS protein with base mispairs in DNA has been reported in detail (Jiricny, J. et al., 1988, *Nuc. Acids Res.* 16, 7843-7853, which is hereby incorporated herein by reference). Besides simplicity, a major advantage of this method for detecting the DNA:protein complex over other suitable methods is the practical lack of a limitation on the size of DNA molecules that can be detected in DNA:protein duplexes. Therefore, this embodiment of this method is useful for detecting single base sequence differences between DNA fragments as large as can be practically handled without shearing, at least 50 kbp.

Another suitable analytical method for detecting the DNA:protein complex between the mispair recognition protein and a cognate mispair in a DNA duplex comprises the steps of separating the DNA:protein complexes from DNA that does not form such complexes on the basis of electrophoretic mobility, and detecting the DNA in the less mobile DNA:protein complexes. The DNA in the DNA:protein complexes may be detected by any of the usual standard means for detection of DNA in gel electrophoresis, including staining with dyes or annealing with complementary strands of radioactive DNA. Detecting complexes comprising the MutS base mispair recognition protein and mispairs in DNA duplexes is also described in the foregoing reference (Jiricny, J. et al., 1988, *Nuc. Acids Res.*, 16, 7843-7853). Under the usual conditions employed in the art for detecting specific DNA:protein complexes by gel electrophoresis,

12

complex formation of a protein with a double-stranded DNA fragment of up to several hundred base pairs is known to produce distinguishable mobility differences.

Other suitable analytical methods for detecting the DNA:MutS protein complex include immunodetection methods using an antibody specific for the base mispair recognition protein. For example, antibodies specific for the *E. coli* MutS protein have been prepared readily by standard immunological techniques. Accordingly, one immunodetection method for complexes of MutS protein with DNA comprises the steps of separating the DNA:protein complexes from DNA that does not form such complexes by immunoprecipitation with an antibody specific for MutS protein, and detecting the DNA in the precipitate. According to the practice of this aspect of this invention, quantitative immunoassay methods known in the art may be employed to determine the number of single base mispairs in homologous regions of two DNA molecules, based upon calibration curves that can be established using complexes of a given mispair recognition protein with DNA duplexes having known numbers of mispairs.

In addition to methods that merely detect base sequence differences, this invention further relates to a method for both detecting and localizing individual base sequence differences within homologous regions of two DNA molecules comprising the steps of contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting double-stranded DNA duplexes with a protein that recognizes at least one base mispair under conditions such that the protein forms specific complexes with its cognate mispairs and thereby directs modification of at least one strand of the DNA in the resulting DNA:protein complexes in the vicinity of the DNA:protein complex, and determination of the location of the resulting DNA modification relative to a known sequence within the homologous regions of the DNAs by a suitable analytical method.

In the method of the present invention for localization of single base differences, there is provided a suitable means for modifying at least one strand of the DNA duplex in the vicinity of the bound mispair recognition protein. The modification may be any alteration for which there is a means of detection, for instance a chemical modification including breaking of a chemical bond resulting in, as examples, cleavage between nucleotides of at least one DNA strand or removal of a base from the sugar residue of a nucleotide. Specific means for modifying DNAs in the vicinity of the DNA:protein complex are provided below for several embodiments of this aspect of the invention, together with interpretations of the phrase "in the vicinity of", as appropriate to the practical limitations of the modification approach in each instance.

In the practice of one embodiment of this aspect of this invention, the mispair recognition protein is the product of the *mutS* gene of *E. coli* or is another functionally homologous protein; and the step in which the DNA is modified in the vicinity of the DNA:protein complex further comprises contacting the DNA:MutS protein complex with a defined *E. coli* DNA mismatch correction system under conditions such that single-stranded gaps are produced in the vicinity of the complexed protein.

The complete defined mismatch correction system comprises the following purified components: *E. coli* MutH, MutL, and MutS proteins, DNA helicase II, single-strand DNA binding protein, DNA polymerase III holoenzyme,

5,459,039

13

exonuclease I, DNA ligase, ATP, and the four deoxynucleoside triphosphates. This set of proteins can process seven of the eight base-base mismatches in a strand-specific reaction that is directed by the state of methylation of a single GATC sequence located 1 kilobase from the mispair. This defined system is described further in Example 1, below. It should be noted that the lack of ability to repair C-C base mispairs in this embodiment of this aspect of the present invention is not a major limitation of the method for detecting all possible base sequence differences between any two naturally occurring DNA sequences because mutations apparently due to C-C mispairing during DNA replication appear arise most infrequently in vivo.

For the purposes of generating single-stranded gaps in the vicinity of the DNA:MutS protein complexes, DNA duplexes containing mispaired base pairs are contacted with the defined mismatch correction system under the standard conditions described in Example 1, Table 3 (Complete reaction), except for the following differences: exogenous dNTP's are omitted or, preferably, 2',3'-dideoxynucleoside-5'-triphosphates (ddNTPs) are added at 100 uM with dNTPs at 10 uM, to inhibit repair of single-strand gaps; and DNA ligase may be omitted from the reaction. The requirement for methyl-directed strand incision by MutH may be obviated by provision of a single-strand nick by some other means within the vicinity of the mispair, as described in Example 1, FIG. 5. A suitable means for inducing such nicks in unmethylated DNA is limited contact with a nuclease, DNase I, for example; under conditions that are well known in the art, this approach creates nicks randomly throughout double-stranded DNA molecules at suitable intervals for allowing the mispair correction system to create single-stranded gaps in the vicinity of a mispair anywhere in the DNA.

It should be noted that in this embodiment of this method for localizing mismatch base pairs, "in the vicinity of" a base mispair is defined practically by the size of the single-strand gaps typically observed under the above conditions, namely up to about one kbp from the mismatch base pair. Further, in this embodiment, the method for determining the locations of these single-stranded gaps within the DNA duplex comprises the steps of cleaving the DNA with a single-strand specific endonuclease and at least one restriction endonuclease, and comparing the electrophoretic mobilities of the resulting modified DNA fragments with DNA restriction fragments not contacted with the defined mismatch correction system. Suitable single-strand specific endonucleases include the S1 single-strand specific nuclease, for example, or other functionally similar nucleases well known in the art. Additional restriction mapping may be performed as needed to further localize any fragment modifications observed in initial applications of the method, until, if desired, a restriction fragment of convenient size for direct sequence determination is obtained for direct comparisons of sequences of the two DNA molecules in the vicinity of the base sequence difference.

The present invention further relates, in part, to forms of mispair recognition proteins which have been altered to provide an inherent means for modifying at least one strand of the DNA duplex in the vicinity of the bound mispair recognition protein.

In a principal embodiment of this aspect of this invention, the altered mispair recognition protein is the modified product of the mutS gene of *E. coli* or is another functionally homologous modified protein to which is attached an hydroxyl radical cleaving function; and the DNA modification step in the DNA mispair localization method further

14

comprises contacting this modified protein with the DNA under conditions such that the radical cleaving function cleaves at least one strand of the DNA in the vicinity of the protein.

Several methods for attaching an hydroxyl radical cleaving function to a DNA binding protein are known in the art. For example, lysyl residues may be modified by chemically attaching the 1,10-phenanthroline-copper complex to lysine residues, resulting in conversion of a DNA binding protein into a highly efficient site-specific nuclease that cleaved both DNA strands (in the presence of hydrogen peroxide as a coreactant) within the 20 base pair binding site of the protein, as determined by DNase I footprinting (C. -H. Chen and D. S. Sigman, 1987, *Science*, 237, 1197). Chemical attachment of an EDTA-iron complex to the amino terminus of another DNA binding protein similarly produced a sequence specific DNA cleaving protein that cut both strands of the target DNA within a few bases of recognition site of similar size (J. P. Sluka, et al., 1987, *Science*, 235, 777).

An alternate means for attaching the hydroxyl radical cleaving function to this same protein involved extension of the amino terminus with the three amino acids, Gly-Gly-His, which is consensus sequence for the copper-binding domain of serum albumin (D. P. Mack et al., 1988, *J. Am. Chem. Soc.*, 110, 7572-7574). This approach allows for preparation of such an artificial DNA cleaving protein directly by recombinant methods, or by direct synthesis using standard solid phase methods, when the peptide is sufficiently short as it was in this case (55 residues including the 3 added amino acids), thereby avoiding the need for an additional chemical modification step of the reagent which is both time consuming and difficult in large scale production. In contrast to the EDTA-iron complex, the particular peptide sequence constructed in this instance cleaved only one example out of four recognition sites in different sequence environments.

Nevertheless, one skilled in the art of protein engineering would appreciate that this general approach for converting a DNA binding protein into a DNA cleaving protein by attachment of an hydrogen radical cleavage function is widely applicable. Hence, DNA base mispair recognition proteins which normally only bind to DNA are modified to cleave DNA by attachment of an hydroxyl radical cleavage function, according to the practice of this aspect of this invention, without undue experimentation, by adjustment of appropriate variables taught in the art, particularly the chemical nature and length of the "spacer" between the protein and the metal binding site.

In the DNA sequence localization method according to this embodiment which employs a modified DNA base mispair recognition protein with attached hydroxyl radical cleavage function, the means for modification of the DNA:protein complex is a suitable metal ion and associated cofactor or cofactors, and the modification comprises double-stranded cleavage of the DNA within the vicinity of any cognate base mispair wherein the "vicinity" substantially corresponds to the sequence of DNA protected by the binding of the protein to a base mispair, generally within about 20 base pairs. A single-strand specific nuclease, S1, for instance, may be used to augment cleavage by the modified base mispair recognition protein in the event that a single-strand bias is suspected in the cleavage of any DNAs with which the protein forms a specific complex.

Additional altered forms of mispair recognition proteins that modify at least one strand of the DNA in a DNA:protein complex in the vicinity of the bound protein according to the present invention include proteins comprising the portions

5,459,039

15

or "domains" of the unmodified base mispair recognition enzymes that are essential for binding to a DNA mispair. These essential domains comprise peptides in the unmodified protein which are made resistant to proteolytic digestion by formation of specific DNA:protein complexes at cognate DNA base mispairs. These essential DNA binding domains further comprise peptide sequences that are most highly conserved during evolution; such conserved domains are evident, for example, in comparisons of the sequences of the *E. coli* MutS protein with functionally homologous proteins in *S. typhimurium* and other structurally similar proteins. Accordingly, peptide sequences of a DNA base mispair recognition protein that are protected from proteases by formation of specific complexes with mispairs in DNA and, in addition or in the alternative, are evolutionarily conserved, form the basis for a particularly preferred embodiment of this aspect of the present invention, since such peptides constitute less than half the mass of the intact protein and, therefore, are advantageous for production and, if necessary, for chemical modification to attach a cleavage function for conversion of the DNA binding protein into a DNA cleavage protein specific for sites of DNA base mispairs.

The present invention also comprises another *E. coli* DNA mispair recognition protein that recognizes only A-G mispairs without any requirement for hemimethylation. This protein, the product of the *mutY* gene, is a glycosylase which specifically removes the adenine from an A-G mispair in a DNA duplex. The Mut Y protein has been purified to near homogeneity by virtue of its ability to restore A-G to C-G mismatch correction to cell-free extracts (K. G. Au et al., *Proc. Nat. Acad. Sci. U.S.A.*, 85, 9163, 1988) of a *mutS* *mutY* double mutant strain of *E. coli*, as described in Example 2, below. It is a 36 kDa polypeptide that apparently exists as a monomer in solution. MutY, an AP endonuclease, DNA polymerase I, and DNA ligase are sufficient to reconstitute MutY-dependent, A-G to C-G repair in vitro. A DNA strand that has been depurinated thusly by the MutY protein is susceptible to cleavage by any of several types of AP (apurinic) endonuclease (e.g., human AP endonuclease II) or by piperidine, under conditions that are well known in the art. The cleavage products are then analyzed by gel electrophoresis as in DNA sequencing. Accordingly, this MutY protein is useful in a method for the specific detection and localization of A-G mispairs, according to the practice of the present invention.

The full novelty and utility of the present invention may be further appreciated by reference to the following brief description of selected specific embodiments which advantageously employ various preferred forms of the invention as applied to a common problem in genetic mapping of point mutations in the human genome. In the course of constructing gene linkage maps, for example, it is frequently desirable to compare the sequence of a DNA cloned fragment comprising twenty or more kbp of unknown sequence (except, perhaps, for a few restriction enzyme recognition sites) with homologous sequences in DNA extracted from a human tissue sample. While fragments containing sequences homologous to the cloned DNA fragment can be detected in the human tissue DNA by the well known "Southern" blotting method using radiolabeled DNA of the clone, as explained in the Background section, detection and localization of all the sequence differences between such a clone and a human DNA sample would be a long and arduous task at best using the best methods available in the prior art, including restriction enzyme mapping and direct DNA sequencing.

16

In contrast, substantially all base pairs in the entire homologous sequence of the cloned DNA fragment are compared to those of the human tissue DNA, most advantageously in a single test according to the present invention, merely by contacting both strands of the human tissue DNA molecule with both radiolabeled complementary strands of the second DNA molecule (usually without separation from the cloning vector DNA) under conditions such that base pairing occurs, contacting the resulting DNA duplexes with the *E. coli* MutS protein that recognizes substantially all base pair mismatches under conditions such that the protein forms specific complexes with its cognate mispairs, and detecting the resulting DNA:protein complexes by contacting the complexes with a membranous nitrocellulose filter under conditions such that protein:DNA complexes are retained while DNA not complexed with protein is not retained, and measuring the amount of DNA in the retained complexes by standard radiological methods.

If the above detection test indicates the presence of sequence differences between the human tissue DNA and the cloned DNA and localization is required, or, in the alternative, if such differences are suspected and localization as well as detection of them is desired in a first analysis, the another method of this invention may be applied for these purposes. An embodiment of this aspect of the invention that may be most advantageously employed comprises the steps of contacting both strands of the human tissue DNA molecule with both radiolabeled complementary strands of the second DNA molecule (usually without separation from the cloning vector DNA) under conditions such that base pairing occurs, contacting the resulting DNA duplexes with a modified form of MutS protein of *E. coli* to which is attached an hydroxyl radical cleaving function under conditions such that the radical cleaving function cleaves both strands of the DNA within about 20 base pairs of substantially all DNA base mispairs. In the absence of any DNA base mispairs in the DNA duplexes comprising complementary strands of the human tissue and cloned DNAs, no DNA fragments smaller than the cloned DNA (plus vector DNA, if still attached) would be detected. Determination of the location of any double-stranded DNA cleavages by the modified MutS protein to within a few kbp or less of some restriction enzyme cleavage site within the cloned DNA is determined by standard restriction enzyme mapping approaches. If greater precision in localization and identification of a single base difference is desired, sequencing could be confined to those particular fragments of cloned DNA that span at least one base sequence difference localized by this method and are cleaved by a restriction enzyme at the most convenient distance of those sequence differences for direct sequencing.

The following Examples are provided for further illustrating various aspects and embodiments of the present invention and are in no way intended to be limiting of the scope.

EXAMPLE 1

DNA Mismatch Correction in a Defined System

In order to address the biochemistry of methyl-directed mismatch correction, the reaction has been assayed in vitro using the type of substrate illustrated in FIG. 1. Application of this method to cell-free extracts of *E. coli* (A. -L. Lu, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983) confirmed in vivo findings that methyl-directed repair requires the products of four mutator genes, *mutH*, *mutL*, *mutS* and *uvrD* (also called *mutU*), and also demonstrated a

5,459,039

17

requirement for the *E. coli* single-strand DNA binding protein (SSB). The dependence of in vitro correction on mutH, mutL, and mutS gene products has permitted isolation of these proteins in near homogeneous, biologically active forms. The 97-kD MutS protein binds to mismatched DNA base pairs; the 70-kD MutL protein binds to the MutS-Heteroduplex complex (M. Grilley, K. M. Welsh, S. -S. Su, P. Modrich, *J. Biol. Chem.* 264, 1000, 1989); and the 25-kD MutH protein possesses a latent endonuclease that incises the unmethylated strand of a hemimethylated d(GATC) site (K. M. Welsh, A. -L. Lu, S. Clark, P. Modrich, *J. Biol. Chem.* 262, 15624, 1987), with activation of this activity depending on interaction of MutS and MutL with a heteroduplex in the presence of ATP (P. Modrich, *J. Biol. Chem.* 264, 6597, 1989). However, these three Mut proteins together with SSB and the DNA helicase II product of the *uvrD* (*mutU*) gene (I. D. Hickson, H. M. Arthur, D. Bramhill, P. T. Emmerson, *Mol. Gen. Genet.* 190, 265, 1983) are not sufficient to mediate methyl-directed repair. Below is described identification of the remaining required components and reconstitution of the reaction in a defined system.

Protein and cofactor requirements for mismatch correction. Methyl-directed mismatch correction occurs by an excision repair reaction in which as much as several kilobases of the unmethylated DNA strand is excised and resynthesized (A. -L. Lu, K. Welsh, S. Clark, S. -S. Su, P. Modrich, *Cold Spring Harbor Symp. Quant. Biol.* 49, 589, 1984). DNA polymerase I, an enzyme that functions in a number of DNA repair pathways, does not contribute in a major way to methyl-directed correction since extracts from a *polA* deletion strain exhibit normal levels of activity. However extracts derived from a *dnaZ*^{ts} strain are temperature sensitive for methyl-directed repair in vitro (Table 1).

TABLE 1

Requirement for τ and γ Subunits of DNA Polymerase III Holoenzyme in Mismatch Repair				
Extract	DNA Pol III addition	Mismatch Correction (fmol/h/mg)		Activity ratio
genotype	(ng)	42°	34°	42°/34°
<i>dnaZ</i> ^{ts}	—	8	91	0.09
	57 ng	75	160	0.47
<i>dnaZ</i> ⁺	—	150	160	0.94
	57 ng	160	160	1.0

Extracts from strains AX727 (*lac thi str^R dnaZ20-16*) and AX729 (as AX727 except *purE dnaZ*⁺) were prepared as described (A. -L. Lu, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983). Samples (110 μ g of protein) were mixed with 0.8 μ l of 1M KCl and water to yield a volume of 7.2 μ l, and preincubated at 42° or 34° C. for 2.5 minutes. All heated samples were then placed at 34° C. and supplemented with 2.2 μ l of a solution containing 0.1 μ g (24 fmol) of hemimethylated G-T heteroduplex DNA, 16 ng of MutL protein, 50 ng of MutS protein, and buffer and nucleotide components of the mismatch correction assay (A. -L. Lu, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983). DNA polymerase III holoenzyme (57 ng in 0.6 μ l) or enzyme buffer was then added, and incubation at 34° C. was continued for 60 min. Heated extracts were supplemented with purified MutL and MutS proteins because these components are labile at 42° C. Activity measurements reflect the correction of heteroduplex sites.

The *dnaZ* gene encodes the T and γ subunits of DNA polymerase III holoenzyme (M. Kodaira, S. B. Biswas, A. Kornberg, *Mol. Gen. Genet.* 192, 80, 1983; D. A. Mullin, C. L. Woldringh, J. M. Henson, J. R. Walker, *Mol. Gen. Genet.* 192, 73, 1983), and mismatch correction activity is largely restored to heated extracts of the temperature-sensitive mutant strain by addition of purified polymerase III holoenzyme. Since DNA polymerase III holoenzyme is highly processive, incorporating thousands of nucleotides per DNA

18

binding event, the involvement of this activity is consistent with the large repair tracts associated with the methyl-directed reaction.

Additional data indicate that purified MutH, MutL, and MutS proteins, DNA helicase II, SSB, and DNA polymerase III holoenzyme support methyl-directed mismatch correction, but this reaction is inhibited by DNA ligase, an enzyme that is shown below to be required to restore covalent continuity to the repaired strand. This observation led to isolation of a 55-kD stimulatory protein that obviates ligase inhibition. The molecular weight and N-terminal sequence of this protein indicated identity to exonuclease I (G. J. Phillips and S. R. Kushner, *J. Biol. Chem.* 262, 455, 1987), and homogeneous exonuclease I readily substitutes for the 55-kD stimulatory activity (Table 2). Thus, exonuclease I and the six activities mentioned above mediate efficient methyl-directed mismatch correction in the presence of ligase to yield product molecules in which both DNA strands are covalently continuous.

TABLE 2

Stimulation of in vitro Methyl- Directed Correction by Exonuclease I.	
Protein added	Mismatch correction (fmol/20 min)
None	1
55-kD protein	18
Exonuclease I	18

Reactions (10 μ l) contained 0.05M HEPES (potassium salt, pH 8.0), 0.02M KCl, 6 mM MgCl₂, bovine serum albumin (0.05 mg/ml), 1 mM dithiothreitol, 2 mM ATP, 100 μ M (each) dATP, dCTP, dGTP, and dTTP, 25 μ M β -AND⁺, 0.1 μ g of hemimethylated, covalently closed G-T heteroduplex DNA (FIG. 1, methylation on c strand, 24 fmol), 0.26 ng of MutH (K. M. Welsh, A. -L. Lu, S. Clark, P. Modrich, *J. Biol. Chem.* 262, 15624, 1987), 17 ng of MutL (M. Grilley, K. M. Welsh, S. -S. Su, P. Modrich, *J. Biol. Chem.* 264, 1000, 1989), 35 ng of MutS (S. -S. Su and P. Modrich, *Proc. Natl. Acad. Sci. USA* 83, 5057, 1986), 200 ng of SSB (T. M. Lohman, J. M. Green, R. S. Beyer, *Biochemistry* 25, 21, 1986; U.S. Biochemical Corp.), 10 ng of DNA helicase II (K. Kumura and M. Sekiguchi, *J. Biol. Chem.* 259, 1560, 1984), 20 ng of *E. coli* DNA ligase (U.S. Biochemical Corp.), 95 ng of DNA polymerase III holoenzyme (C. McHenry and A. Kornberg, *J. Biol. Chem.* 252, 6478, 1977), and 1 ng of 55-kD protein or exonuclease I (U.S. Biochemical Corp.) as indicated. Reactions were incubated at 37° C. for 20 minutes, quenched at 55° C. for 10 minutes, chilled on ice, and then digested with Xho I or Hind III endonuclease to monitor correction. Repair of the G-T mismatch yielded a only the G-C containing, Xho I-sensitive product.

The requirements for repair of a covalently closed G-T heteroduplex (FIG. 1) are summarized in Table 3 (Closed circular). No detectable repair was observed in the absence of MutH, MutL, or MutS proteins or in the absence of DNA polymerase III holoenzyme, and omission of SSB or exonuclease I reduced activity by 85 to 90 per cent.

TABLE 3

Protein and Cofactor Requirements for Mismatch Correction in a Defined System.		
Reaction conditions	Closed Circular Heteroduplex	Open Circular Heteroduplex
Complete	15	17 (No MutH, No ligase)
minus MutH	<1	—
minus MutL	<1	<1
minus MutS	<1	<1
minus DNA polymerase III holoenzyme	<1	<1

5,459,039

19

TABLE 3-continued

Reaction conditions	Mismatch correction (fmol/20 min)	
	Closed Circular Heteroduplex	Open Circular Heteroduplex
minus SSB	2	1.4
minus exonuclease I	2	<1
minus DNA helicase II	16	15
minus helicase II, plus immune serum	<1	<1
minus helicase II, plus pre-immune serum	14	NT
minus Ligase/AND ⁺	14	—
minus MgCl ₂	<1	NT
minus ATP	<1	NT
minus dNTP's	<1	NT

Reactions utilizing covalently closed G-T heteroduplex (modification on c strand) were performed as described in the legend to TABLE 2 except that 1.8 ng of exonuclease I was used. Repair of open circular DNA was performed in a similar manner except that MutH, DNA ligase, and β -AND⁺ were omitted from all reactions, and the hemimethylated G-T heteroduplex (modification on c strand) had been incised with MutH protein as described in the legend to FIG. 4. When present, rabbit antiserum to helicase II or pre-immune serum (5 μ g protein) was incubated at 0° C. for 20 minutes with reaction mixtures lacking MgCl₂; the cofactor was then added and the assay was performed as above. Although not shown, antiserum inhibition was reversed by the subsequent addition of more helicase II. With the exception of the DNA polymerase III preparation, which contained about 15% by weight DNA helicase II (text), the purity of individual protein fractions was $\geq 95\%$. NT -- not tested.

These findings are in accord with previous conclusions concerning requirements of the methyl-directed reaction. However, in contrast to observations *in vivo* and in crude extracts indicating a requirement for the uvrD product, the reconstituted reaction proceeded readily in the absence of the added DNA helicase II (Table 2). Nevertheless, the reaction was abolished by antiserum to homogeneous helicase II, suggesting a requirement for this activity and that it might be present as a contaminant in one of the other proteins. Analysis of these preparations for their ability to restore mismatch repair to an extract derived from a uvrD (mutU) mutant and for the physical presence of helicase II by immunoblot assay revealed that the DNA polymerase III holoenzyme preparation contained sufficient helicase II (13 to 15 per cent of total protein by weight) to account for the levels of mismatch correction observed in the defined system. Similar results were obtained with holoenzyme preparations obtained from two other laboratories. The purified system therefore requires all the proteins that have been previously implicated in methyl-directed repair.

The rate of correction of the closed circular heteroduplex was unaffected by omission of DNA ligase (Table 3), but the presence of this activity results in production of a covalently closed product. Incubation of a hemimethylated, supercoiled G-T heteroduplex with all seven proteins required for correction in the presence of DNA ligase resulted in extensive formation of covalently closed, relaxed, circular molecules. Production of the relaxed DNA was dependent on MutS (FIG. 2) and MutL proteins, and the generation of this species was associated with heteroduplex repair (FIG. 2). Correction also occurred in the absence of ligase, but in this case repair products were open circular molecules, the formation of which depended on the presence of MutS (FIG. 2). Since MutS has no known endonuclease activity but does recognize mismatches, it is inferred that open circular molecules are the immediate product of a mismatch-provoked

20

excision repair process. Ligase closure of the strand break(s) present in this species would yield the covalently closed, relaxed circular product observed with the complete system.

The set of purified activities identified here as being important in methyl-directed repair support efficient correction. In the experiments summarized in Table 3, the individual proteins were used at the concentrations estimated to be present in the standard crude extract assay for correction as calculated from known specific activity determinations. Under such conditions the rate and extent of mismatch repair in the purified system are essentially identical to those observed in cell-free extracts.

DNA sites involved in repair by the purified system. The single d(GATC) sequence within the G-T heteroduplex shown in FIG. 1 is located 1024 base pairs from the mismatch. Despite the distance separating these two sites, correction of the mismatch by the purified system responded to the state of modification of the d(GATC) sequence as well as its presence within the heteroduplex (FIG. 3). A substrate bearing d(GATC) methylation on both DNA strands did not support mismatch repair nor did a related heteroduplex in which the d(GATC) sequence was replaced by d(GATT). However, each of the two hemimethylated heteroduplexes were subject to strand-specific correction, with repair in each case being restricted to the unmodified DNA strand. With a heteroduplex in which neither strand was methylated, some molecules were corrected on one strand, and some were corrected on the other. As can be seen, the hemimethylated heteroduplex bearing methylation on the complementary DNA strand was a better substrate than the alternative configuration in which modification was on the viral strand, with a similar preference for repair of the viral strand being evident with the substrate that was unmethylated on either strand. This set of responses of the purified system to the presence and state of modification of d(GATC) sites reproduce effects previously documented *in vivo* and in crude extract experiments (R. S. Lahue, S. -S. Su, P. Modrich, *Proc. Natl. Acad. Sci. USA* 84, 1482, 1987).

The efficiency of repair by the methyl-directed pathway depends not only on the nature of the mismatch, but also on the sequence environment in which the mismatch is embedded (P. Modrich, *Ann. Rev. Biochem.* 56, 435, 1987). To assess the mismatch specificity of the purified system under conditions where sequence effects are minimized, a set of heteroduplexes were used in which the location and immediate sequence environment of each mismatch are essentially identical (S. -S. Su, R. S. Lahue, K. G. Au, P. Modrich, *J. Biol. Chem.* 263, 6829, 1988). This analysis (Table 4) showed that the

TABLE 4

Correction Efficiencies for Different Mismatches.					
Heteroduplex	Markers	Methylation State			
		C ⁺ V ⁻		C ⁻ V ⁺	
		Rate	Bias	Rate	Bias
60 C 5'-CTCGA G AGCTT	Xho I	1.2	>18	0.38	>5
V 3'-GAGCT T TCGAA	Hind III				
C 5'-CTCGA G AGCTG	Xho I	1.1	>17	0.38	>6
V 3'-GAGCT G TCGAC	Pvu II				
C 5'-ATCGA T AGCTT	Cla I	1.0	>16	0.24	3
V 3'-TAGCT T TCGAA	Hind III				
C 5'-ATCGA A AGCTT	Hind III	0.88	>20	0.20	>7
V 3'-TAGCT A TCGAA	Cla I				
65 C 5'-CTCGA A AGCTT	Hind III	0.61	17	0.28	>5

5,459,039

21

TABLE 4-continued

Correction Efficiencies for Different Mismatches.					
Heteroduplex	Markers	Methylation State			
		C ⁺ V ⁻		C ⁻ V ⁺	
		Rate	Bias	Rate	Bias
V 3'-GAGCT C TCGAA	Xho I				
C 5'-GTCGA C AGCTT	Sal I	0.60	12	0.23	>4
V 3'-CAGCT T TCGAA	Hind III				
C 5'-GTCGA A AGCTT	Hind III	0.44	>13	0.21	5
V 3'-CAGCT G TCGAA	Sal I				
C 5'-CTCGA C AGCTG	Pvu II	0.04	NS	<0.04	NS
V 3'-GAGCT C TCGAC	Xho I				

Correction of the eight possible base-base mispairs was tested with the set of covalently closed heteroduplexes described previously including the G-T substrate shown in FIG. 1. With the exception of the mispair and the variations shown at the fifth position on either side, all heteroduplexes were identical in sequence. Each DNA was tested in both hemimethylated configurations under complete reaction conditions (Table 3, closed circular heteroduplex) except that samples were removed at 5-minute intervals over a 20 minute period in order to obtain initial rates (fmol/min). c and v refer to complementary and viral DNA strands, and Bias indicates the relative efficiency of mismatch repair occurring on the two DNA strands (ratio of unmethylated to methylated) as determined 60 minutes after the reaction was started. NS - not significant. With the exception of the C-C heteroduplexes, repair in the absence of MutS protein was less than 20% (in most cases <10%) of that observed in its presence (not shown).

purified system is able to recognize and repair in a methyl-directed manner seven of the eight possible base-base mismatches, with C-C being the only mispair that was not subject to significant correction. Table 3 also shows that the seven corrected mismatches were not repaired with equal efficiency and that in the case of each heteroduplex, the hemimethylated configuration modified on the complementary DNA strand was a better substrate than the other configuration in which the methyl group was on the viral strand. These findings are in good agreement with patterns of repair observed with this set of heteroduplexes in *E. coli* extracts (Although the patterns of substrate activity observed in extracts and in the purified system are qualitatively identical, the magnitude of variation observed differs for the two systems. Hemimethylated heteroduplexes modified on the complementary DNA strand are better substrates in both systems, but in extracts such molecules are repaired at about twice the rate of molecules methylated on the viral strand. In the purified system these relative rates differ by factors of 2 to 4. A similar effect may also exist with respect to mismatch preference within a given hemimethylated family. Although neither system repairs C-C, the rates of repair of other mismatches vary by a factors of 1.5 to 2 in extracts but by factors of 2 to 3 in the defined system.).

Strand-specific repair directed by a DNA strand break. Early experiments on methyl-directed repair in *E. coli* extracts led to the proposal that the strand-specificity of the reaction resulted from endonucleolytic incision of an unmethylated DNA strand at a d(GATC) sequence. This idea was supported by the finding that purified MutH protein has an associated, but extremely weak d(GATC) endonuclease that is activated in a mismatch-dependent manner in a reaction requiring MutL, MutS, and ATP. The purified system has been used to explore this effect more completely.

The two hemimethylated forms of the G-T heteroduplex shown in FIG. 1 were incised using high concentrations of purified MutH protein to cleave the unmethylated DNA strand at the d(GATC) sequence (>>pGpApTpC). After removal of the protein, these open circular heteroduplexes were tested as substrates for the purified system in the absence of DNA ligase. Both open circular species were

22

corrected in a strand-specific manner and at rates similar to those for the corresponding covalently closed heteroduplexes (FIG. 4). As observed with closed circular heteroduplexes, repair of the MutH-cleaved molecules required MutL, MutS, SSB, DNA polymerase III holoenzyme, and DNA helicase II (FIG. 4 and open circle entries of Table 2), but in contrast to the behavior of the closed circular substrates, repair of the mismatch within the open circular molecules occurred readily in the absence of MutH protein. Thus prior incision of the unmethylated strand of a d(GATC) site can bypass the requirement for MutH protein in strand-specific mismatch correction.

The nature of the MutH-independent repair was examined further to assess the effect of ligase on the reaction and to determine whether a strand break at a sequence other than d(GATC) can direct correction in the absence of MutH protein (FIG. 5). As mentioned above, a covalently closed G-T heteroduplex that lacks a d(GATC) sequence is not subject to repair by the purified system in the presence (FIG. 3) or absence of DNA ligase. However, the presence of one strand-specific, site-specific break is sufficient to render this heteroduplex a substrate for the purified system in the absence of ligase and MutH protein (FIG. 5). Repair of this open circular heteroduplex was limited to the incised, complementary DNA strand, required presence of MutL and MutS proteins, DNA polymerase III, and SSB, and correction of the molecule was as efficient as that observed with the hemimethylated heteroduplex that had been cleaved by MutH at the d(GATC) sequence within the complementary strand. Although the presence of a strand break is sufficient to permit strand-specific correction of a heteroduplex in the absence of MutH and ligase, the presence of the latter activity inhibited repair not only on the heteroduplex lacking a d(GATC) sequence but also on both hemimethylated molecules that had been previously incised with MutH protein (FIG. 5). This inhibition by ligase was circumvented by the presence of MutH protein, but only if the substrate contained a d(GATC) sequence, with this effect being demonstrable when both types of heteroduplex were present in the same reaction (FIG. 5, last column). This finding proves that MutH protein recognizes d(GATC) sites and is consistent with the view that the function of this protein in mismatch correction is the incision of the unmethylated strand at this sequence.

EXAMPLE 2

Purification of MutY Protein

Purification of MutY Protein. *E. coli* RK1517 was grown at 37° C. in 170 liters of L broth containing 2.5 mM KH₂PO₄, 7.5 mM Na₂HPO₄ (culture pH=7.4) and 1% glucose. The culture was grown to an A590 of 4, chilled to 10° C. and cells were harvested by continuous flow centrifugation. Cell paste was stored at -70° C. A summary of the MutY purification is presented in Table 1. Fractionation procedures were performed at 0°-4° C., centrifugation was at 13,000×g, and glycerol concentrations are expressed as volume percent.

Frozen cell paste (290 g) was thawed at 4° C., resuspended in 900 ml of 0.05M Tris-HCl (pH 7.5), 0.1M NaCl, 1 mM dithiothreitol, 0.1 mM EDTA, and cells were disrupted by sonication. After clarification by centrifugation for 1 hr, the lysate (Fraction I, 970 ml) was treated with 185 ml of 25% streptomycin sulfate (wt/vol in 0.05M Tris-HCl (pH 7.5), 0.1M NaCl, 1 mM dithiothreitol, 0.1 mM EDTA) which was added slowly with stirring. After 30 min of

5,459,039

23

additional stirring, the solution was centrifuged for 1 h, and the supernatant (1120 ml) was treated with 252 g of solid ammonium sulfate which was added slowly with stirring. After 30 min of additional stirring, the precipitate was collected by centrifugation for 1 h, resuspended to a final volume of 41 ml in 0.02M potassium phosphate (pH 7.5), 0.1 mM EDTA, 10% (vol/vol) glycerol, 1 mM dithiothreitol, and dialyzed against two 2 l portions of 0.02M potassium phosphate (pH 7.5), 0.1M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol (2 h per change). The dialyzed material was clarified by centrifugation for 10 min to yield Fraction II (45 ml).

Fraction II was diluted 10-fold into 0.02M potassium phosphate (pH 7.5), 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol so that the conductivity of the diluted solution was comparable to that of the dilution buffer containing 0.1M KCl. The dilution was performed on small aliquots of Fraction II, and diluted samples were immediately loaded at 1 ml/min onto a 14.7 cm \times 12.6 cm² phosphocellulose column equilibrated with 0.02M potassium phosphate (pH 7.5), 0.1M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. The column was washed with 400 ml of equilibration buffer, and developed with a 2 liter linear gradient of KCl (0.1 to 1.0M) in 0.02M potassium phosphate (pH 7.5), 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. Fractions containing MutY activity, which eluted at about 0.4M KCl, were pooled (Fraction III, 169 ml).

Fraction III was dialyzed against three 2 liter portions of 5 mM potassium phosphate (pH 7.5), 0.05M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol (2 h per change) until the conductivity was comparable to that of the dialysis buffer. After clarification by centrifugation at for 10 min, the solution was loaded at 0.5 ml/min onto a 21 cm \times 2.84 cm² hydroxylapatite column equilibrated with 5 mM potassium phosphate, pH 7.5, 0.05M KCl, 1 mM dithiothreitol, 10% glycerol. After washing with 130 ml of equilibration buffer, the column was eluted with a 600 ml linear gradient of potassium phosphate (5 mM to 0.4M, pH 7.5) containing 0.05M KCl, 1 mM dithiothreitol, 10% glycerol. Fractions eluting from the column were supplemented with EDTA to 0.1 mM. Peak fractions containing 60% of the total recovered activity, which eluted at about 0.1M potassium phosphate, were pooled (Fraction IV, 24 ml). The remaining side fractions contained impurities which could not be resolved from MutY by MonoS chromatography.

Fraction IV was diluted by addition of an equal volume of 0.05M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. After clarification by centrifugation for 15 min, diluted Fraction IV was loaded at 0.75 ml/min onto a Pharmacia HR 5/5 MonoS FPLC column that was equilibrated with 0.05M sodium phosphate (pH 7.5), 0.1M NaCl, 0.1 mM EDTA, 0.5 mM dithiothreitol, 10% glycerol. The column was washed at 0.5 ml/min with 17 ml of equilibration buffer and developed at 0.5 ml/min with a

TABLE 1

Purification of MutY protein from 290 g of <i>E. coli</i> RK1517				
Fraction	Step	Total Protein mg	Specific Activity units/mg	Yield Percent
I	Extract	10,900	40	(100)
II	Ammonium sulfate	1,350	272	84
III	Phosphocellulose	66	10,800	160
IV	Hydroxylapatite	1.4	136,000	44
V	MonoS	0.16	480,000	18

24

TABLE 1-continued

Purification of MutY protein from 290 g of <i>E. coli</i> RK1517				
Fraction	Step	Total Protein mg	Specific Activity units/mg	Yield Percent

Specific A.G to C-G mismatch correction in cell-free extracts was determined as described previously (Au et al. 1988), except that ATP and glutathione were omitted from the reaction and incubation was for 30 min instead of 1 h. For complementation assays, each 0.01 ml reaction contained RK1517-Y33 extract (mutS mutY) at a concentration of 10 mg/ml protein. One unit of MutY activity is defined as the amount required to convert 1 fmol of A.G mismatch to C-G base pair per h under complementation conditions.

20 ml linear gradient of NaCl (0.1 to 0.4M) in 0.05M sodium phosphate (pH 7.5), 0.1 mM EDTA, 0.5 mM dithiothreitol, 10% glycerol. Fractions with MutY activity, which eluted at approximately 0.2M NaCl, were pooled (Fraction V, 2.6 ml). Fraction V was divided into small aliquots and stored at -70° C.

Assay for MutY-dependent, A.G-specific glycosylase

DNA restriction fragments were labeled at either the 3' or 5' ends with ³²P. Glycosylase activity was then determined in 0.01 ml reactions containing 10 ng end-labeled DNA fragments, 0.02M Tris-HCl, pH 7.6, 1 mM EDTA, 0.05 mg/ml bovine serum albumin, and 2.7 ng MutY. After incubation at 37° C. for 30 min, the reaction mixture was treated with 2.5 \times 10⁻³ units of HeLa AP endonuclease II in the presence of 11 mM MgCl₂ and 0.005% Triton X-100 for 10 min at 37° C. Reactions were quenched by the addition of an equal volume of 80% formamide, 0.025% xylene cyanol, 0.025% bromphenol blue, heated to 80° C. for 2 min, and the products analyzed on an 8% sequencing gel. Control reactions contained either no MutY, no A.G mismatch or no AP endonuclease II.

Strand cleavage at the AP site generated by MutY could also be accomplished by treatment with piperidine instead of treatment with AP endonuclease II. After incubation for 30 min. at 37° C. with MutY as described above, the reaction mixture was precipitated with ethanol in the presence of carrier tRNA, then resuspended in 1M piperidine and heated at 90° C. for 30 min. After two additional ethanol precipitations, changing tubes each time, the pellet was resuspended in a minimum volume of water to which was added an equal volume of 80% formamide, 0.025% xylene cyanol, 0.025% bromphenol blue. The products were then analyzed on an 8% sequencing gel.

For purposes of completing the background description and present disclosure, each of the published articles, patents and patent applications heretofore identified in this specification are hereby incorporated by reference into the specification.

The foregoing invention has been described in some detail for purposes of clarity and understanding. It will also be obvious that various combinations in form and detail can be made without departing from the scope of the invention.

What is claimed is:

1. A method for detecting a base pair mismatch in a DNA duplex, comprising the steps of:

- contacting a DNA duplex potentially containing a base pair mismatch with a protein that recognizes said base pair mismatch under conditions suitable for said protein to form a specific complex only with said DNA duplex having a base pair mismatch, and not with a DNA duplex lacking a base pair mismatch, and
- detecting any said complex as a measure of the presence of a base pair mismatch in said DNA duplex.

5,459,039

25

2. The method of claim 1 wherein said protein is the product of the mutS gene of *Escherichia coli*.

3. The method of claim 2 wherein said protein recognizes all eight possible base pair mismatches.

4. The method of claim 1 wherein said protein is the product of the mutY gene of *Escherichia coli*.

26

5. The method of claim 1 wherein said protein is a homolog of the MutS protein of *Escherichia coli*.

6. The method of claim 1 wherein said protein is a homolog of the MutY protein of *Escherichia coli*.

* * * * *

EXHIBIT B



US005556750A

United States Patent [19][11] **Patent Number:** **5,556,750****Modrich et al.**[45] **Date of Patent:** **Sep. 17, 1996**

[54] **METHODS AND KITS FOR FRACTIONATING A POPULATION OF DNA MOLECULES BASED ON THE PRESENCE OR ABSENCE OF A BASE-PAIR MISMATCH UTILIZING MISMATCH REPAIR SYSTEMS**

[75] Inventors: **Paul L. Modrich**, Chapel Hill, N.C.; **Shin-San Su**, Newton, Mass.; **Karin G. Au**, Durham, N.C.; **Robert S. Lahue**, Northboro; **Deani L. Cooper**, Watertown, both of Mass.; **Leroy Worth, Jr.**, Durham, N.C.

[73] Assignee: **Duke University**, Durham, N.C.

[21] Appl. No.: **145,837**

[22] Filed: **Nov. 1, 1993**

Related U.S. Application Data

[63] Continuation-in-part of Ser. No. 2,529, Jan. 11, 1993, abandoned, which is a continuation of Ser. No. 350,983, May 12, 1989, abandoned.

[51] Int. Cl.⁶ **C12Q 1/68; C12P 19/34**

[52] U.S. Cl. **435/6; 435/91.1; 435/91.2; 435/810; 436/501; 536/22.1; 536/23.1; 536/24.3; 536/24.31; 536/24.32; 536/24.33; 935/77; 935/78; 935/88**

[58] Field of Search **435/6, 91.1, 91.2, 435/810; 436/501; 536/22.1, 23.1, 24.1, 24.3-33; 935/77, 78, 88**

[56] References Cited

U.S. PATENT DOCUMENTS

4,794,075 12/1988 Ford et al. 435/6

FOREIGN PATENT DOCUMENTS

2239456 7/1991 United Kingdom.
9302216 2/1993 WIPO.
9320233 10/1993 WIPO.
9322462 11/1993 WIPO.
9322457 11/1993 WIPO.

OTHER PUBLICATIONS

Wilchek et al. (1988) *Analytical Bioch.*, vol. 171, pp. 1-32.
Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1986, NIH Grant, Abstract (Source: CRISP).
Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1987, NIH Grant, Abstract (Source: CRISP).
Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1988, NIH Grant, Abstract (Source: CRISP).
Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1989, NIH Grant, Abstract (Source: CRISP).
Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1990, NIH Grant, Abstract (Source: CRISP).
Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1991, NIH Grant, Abstract (Source: CRISP).
Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1992, NIH Grant, Abstract (Source: CRISP).
Modrich, "Molecular Mechanisms of DNA-Protein Interaction", 1993, NIH Grant, Abstract (Source: CRISP).
Modrich, "Enzymology of Eukaryotic DNA Mismatch Repair" 1991, NIH Grant, Abstract (Source: CRISP).
Modrich, "Enzymology of Eukaryotic DNA Mismatch Repair" 1992, NIH Grant, Abstract (Source: CRISP).

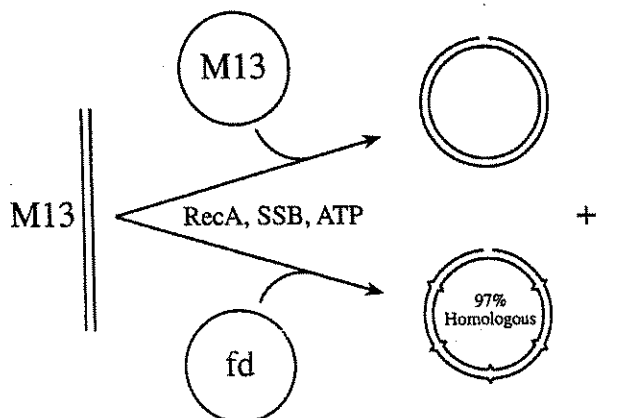
(List continued on next page.)

Primary Examiner—W. Gary Jones
Assistant Examiner—Ardin H. Marschel
Attorney, Agent, or Firm—Lyon & Lyon

[57] ABSTRACT

A diagnostic method for detecting a base pair mismatch in a DNA duplex, comprising the steps of contacting at least one strand of a first DNA molecule with the complementary strand of a second DNA molecule under conditions such that base pairing occurs contacting a DNA duplex potentially containing a base pair mismatch with a mispair recognition protein under conditions suitable for the protein to form a specific complex only with the DNA duplex having a base pair mismatch, and not with a DNA duplex lacking a base pair mismatch, and detecting any complex as a measure of the presence of a base pair mismatch in the DNA duplex.

17 Claims, 7 Drawing Sheets



5,556,750

Page 2

OTHER PUBLICATIONS

- Modrich, "Enzymology of Eukaryotic DNA Mismatch Repair" 1993, NIH Grant, Abstract (Source: CRISP).
- Adams et al. "The Biochemistry of the Nucleic Acids" Chapman & Hall (1986) pp. 221-223.
- Quinones et al. "Expression of the *Escherichia coli* dna Q (mutD) Gene is Inducible" *Mol Gen Genet* (1988) 211:106-112.
- Cotton et al. "Reactivity of Cytosine and Thymine In Single-base-pair Mismatches With Hydroxylamine and Osmium Tetroxide and Its Application to the Study of Mutations" *Proc. Natl. Acad. Sci.* (Jun., 1986) 85:4397-4401.
- Lu et al. "Methyl-directed Repair of DNA Base-pair Mismatches In Vitro" *Proc. Natl. Acad. Sci.* (Aug., 1983) 80:4639-4643.
- Su & Modrich, "*Escherichia coli* mutS-encoded Protein Binds to Mismatched DNA Base Pairs" *Proc. Natl. Acad. Sci.* (Jul., 1986) 83:5057-5061.
- Su et al. "Mispair Specificity of Methyl-directed DNA Mismatch Correction in Vitro" *J.B.C.* (May 15, 1988) 263:6829-6835.
- Jiricny et al. "Mismatch-containing Oligonucleotide Duplexes Bound By The *E.coli* mutS-encoded Protein" *Nucleic Acids Research* (1988) 16:7843-7853.
- Welsh et al. "Isolation and Characterization of the *Escherichia coli* mutL Gene Product" *J.B.C.* (Nov. 15, 1987) 262:15624-15629.
- Au et al. "Initiation of Methyl-directed Mismatch Repair" *J.B.C.* (Jun. 15, 1992) 267:12142-12148.
- Grilley et al. "Isolation and Characterization of the *Escherichia coli* mutL Gene Product" *J.B.C.* (Jan. 15, 1989) 264:1000-1004.
- Su et al. "Gap Formation is Associated With Methyl-Directed Mismatch Correction Under Conditions of Restricted DNA Synthesis" *Genome* (1989) 31:104-111.
- Lahue & Modrich, "DNA Mismatch Correction In A Defined System" *Science* (Jul. 14, 1989) 245:160-164.
- Holmes, Jr. et al. "Strand-specific Mismatch Correction In Nuclear Extracts of Human and *Drosophila* Melanogaster Cell Lines" *Proc. Natl. Acad. Sci.* (Aug. 1990) 87:5837-5841.
- Au et al. "*Escherichia coli* mutY Gene Encodes An Adenine Glycosylase Active on G-A Mispairs" *Proc. Natl. Acad. Sci.* (Nov. 1989) 86:8877-8881.
- Nelson et al. "Genomic Mismatch Scanning A New Approach To Genetic Linkage Mapping" *Nature Genetics* (May, 1993) 4:11-19.
- Au et al. "*Escherichia coli* mutY Gene Product is Required For Specific A-G C.G Mismatch Correction" *Proc. Natl. Acad. Sci.* (Dec. 1988) 85:9163-9166.
- Modrich "Methyl-directed DNA Mismatch Correction" *J.B.C.* (Apr. 25, 1989) 264:6597-6600.
- Lu et al. "Repair of DNA Base-pair Mismatches in Extracts of *Escherichia coli*" *Cold Spring Harbor Laboratory* (1984) Cold Spring Harbor Symposia on Quantitative Biology XLIX:589-596.
- Modrich "DNA Mismatch Correction" *Ann. Rev. Biochem.* (1987) 56:435-466.
- Lahue et al. "Requirements for d(GATC) Sequences in *Escherichia coli* mutHLS Mismatch Correction" *Proc. Natl. Acad. Sci.* (Mar., 1987) 84:1482-1486.
- Lahue & Modrich "Methyl-directed DNA Mismatch Repair in *Escherichia coli*" *Mutation Research* (1988) 198:37-43.
- Modrich "Mechanisms and Biological Effects of Mismatch Repair" *Annu. Rev. Genet.* (1991) 25:229-253.
- Grilley et al. "Mechanisms of DNA-Mismatch Correction" *Mutation Research* (1990) 236:253-267.
- Myers et al. "Detection of Single Base Substitutions by Ribonuclease Cleavage at Mismatches in RNA:DNA Duplexes" *Science* (1985) 230:1242-1246.
- Chen & Sigman "Chemical Conversion of A DNA-Binding Protein Into A Site-Specific Nuclease" *Science* (1987) 237:1197-1201.
- Bianchi & Radding "Insertions, Deletions and Mismatches in Heteroduplex DNA Made By RecA Protein" *Cell* (Dec. 1983) 35:511-520.
- DasGupta & Radding "Polar Branch Migration Promoted By recA Protein: Effect of Mismatched Base Pairs" *Proc. Natl. Acad. Sci.* (Feb. 1982) 79:762-766.
- Rayssiguier et al. "The Barrier To Recombination Between *Escherichia coli* and *Salmonella typhimurium* is Disrupted In Mismatched-Repair Mutants" *Nature* (Nov. 23, 1989) 342:396-401.
- Lu "Influence of GATC Sequences on *Escherichia coli* DNA Mismatch Repair In Vitro" *Journal of Bacteriology* (Mar. 1987) pp. 1254-1259.
- Lu & Chang "A Novel Nucleotide Excision Repair For The Conversion of An A/G Mismatch to C/G Base Pair in *E. coli*" *Cell* (Sep. 9, 1988) 54:805-812.
- Lu & Chang "Repair of Single Base-Pair Transversion Mismatches of *Escherichia coli* In Vitro: Correction of Certain A/G Mismatches Is Independent of dam Methylation and Host mutHLS Gene Functions" *Genetics* (Apr., 1988) 118:593-600.
- Shen & Huang "Effect of Base Pair Mismatches on Recombination Via The RecBCD Pathway" *Mol Gen Genet* (1989) 218:358-360.
- Fang & Modrich "Human Strand-Specific Mismatch Repair Occurs By A Bidirectional Mechanism Similar to That of The Bacterial Reaction" *J.B.C.* (Jun. 5, 1983) 268:11838-11844.
- Hennighausen & Lubon "Interaction of Protein With DNA In Vitro" *Guide to Molecular Cloning Techniques* [Editors: Berger & Kimmel] (1987) 152:721-735.
- Lu and Hsu, "Detection of Single DNA Based Mutations with Mismatch Repair Enzymes," *Genomics* 14:249-255 (1992).
- Marx, J. "DNA Repair Comes Into Its Own," *Science* 266:728-730 (1994).
- Modrich, P., "Mismatch Repair, Genetic Stability, and Cancer," *Science* 266:1959-1960 (1994).
- Priebe et al. "Nucleotide Sequence of the hexA Gene for DNA Mismatch Repair in *Streptococcus pneumoniae* and Homology of hexA to mutS of *Escherichia coli* and *Salmonella typhimurium*," *Journal of Bacteriology* 170:190-196 (1988).
- Reenan & Kolodner, "Isolation and Characterization of Two *Saccharomyces cerevisiae* Genes Encoding Homologs of the Bacterial HexA and MutS Mismatch Repair Proteins" *Genetics* (Dec., 1992) 132:963-973.
- Pang, et al. "Identification and Characterization of the mutL and mutS Gene Products of *Salmonella typhimurium* LT2" *Journal of Bacteriology* (Sep. 1985) 163:1007-1015.
- Grilley, et al. "Bidirectional Excision in Methyl-Directed Mismatch Repair" *J.B.C.* (Jun. 5, 1993) 268:11830-11837.
- Cooper, et al. "Methyl-Directed Mismatch Repair is Bidirectional" *J.B.C.* (Jun. 5, 1993) 268:11823-11829.

U.S. Patent

Sep. 17, 1996

Sheet 1 of 7

5,556,750

V 5'-AAGCTTTCGAG Hind III
C 3'-TTCGAGAGCTC Xho I

FIG. 1.

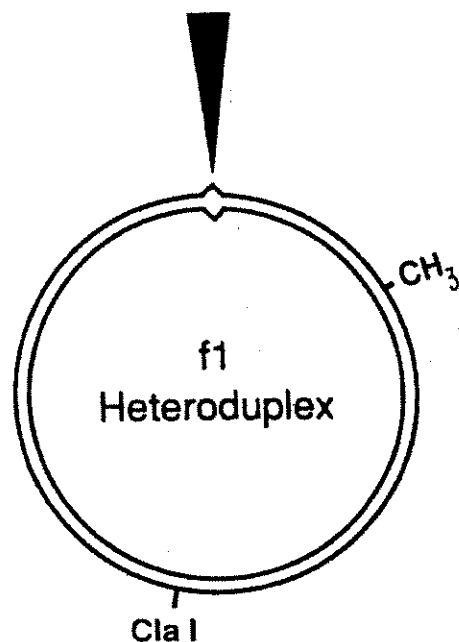
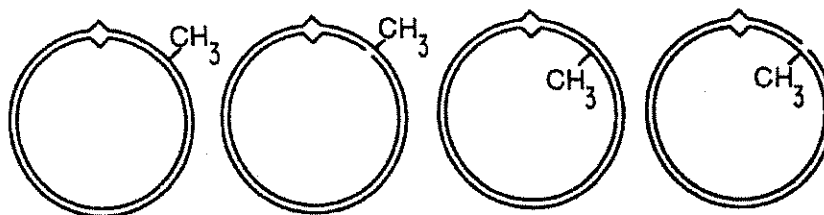


FIG. 4.



Reaction conditions	Repair (fmol/20 min)			
Complete	15 (<1)	17 (<1)	8 (<1)	10 (<1)
- Mut H	<1	18	1	9
- Mut L	<1	<1	<1	<1
- Mut S	<1	<1	<1	1
- SSB	2	<1	<1	<1
- pol III holoenzyme	<1	<1	<1	<1

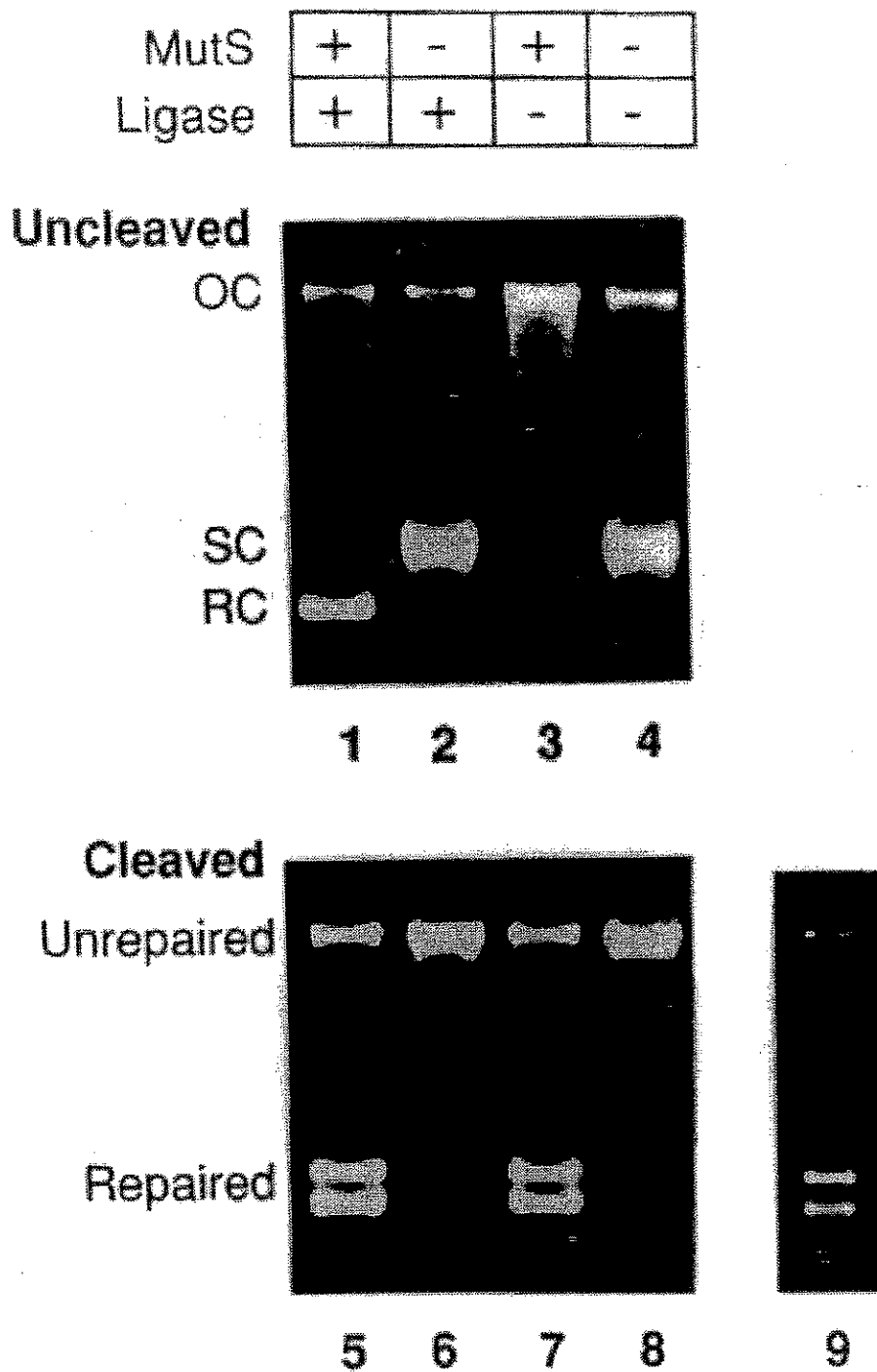
U.S. Patent

Sep. 17, 1996

Sheet 2 of 7

5,556,750

FIG. 2.



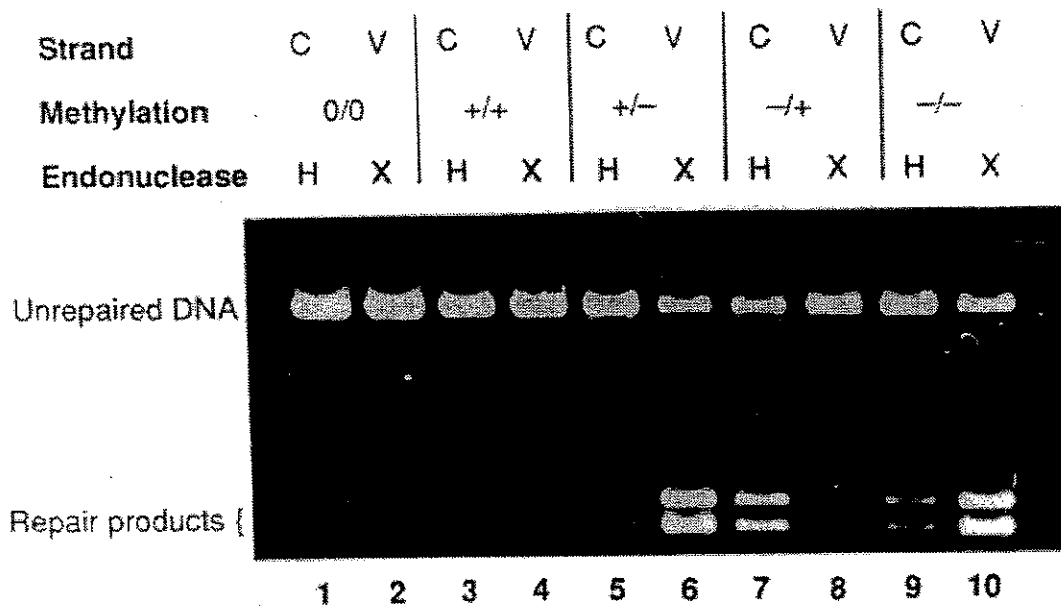
U.S. Patent

Sep. 17, 1996

Sheet 3 of 7

5,556,750

FIG. 3.

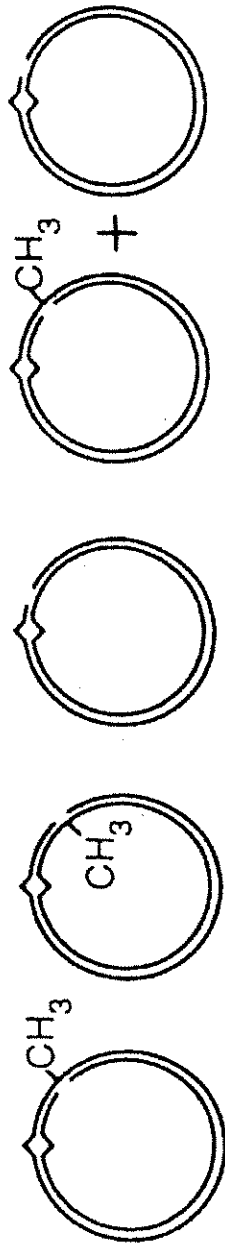


U.S. Patent

Sep. 17, 1996

Sheet 4 of 7

5,556,750



Repair (fmol/20 min)

Ligase MutH

Ligase	MutH	Repair (fmol/20 min)			
		19 (<1)	9 (<1)	11 (<1)	19 (<1)
—	—	19 (<1)	9 (<1)	11 (<1)	9 (<1)
+	—	2	<1	1	2
+	+	20	7	2	15

FIG. 5.

U.S. Patent

Sep. 17, 1996

Sheet 5 of 7

5,556,750

FIG. 6.

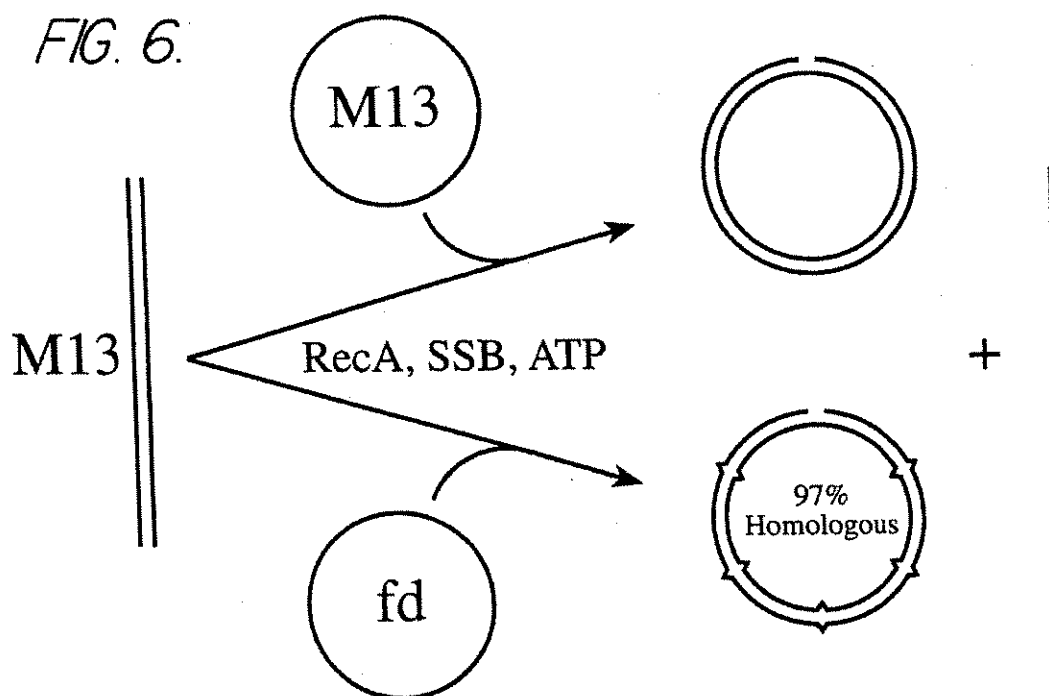
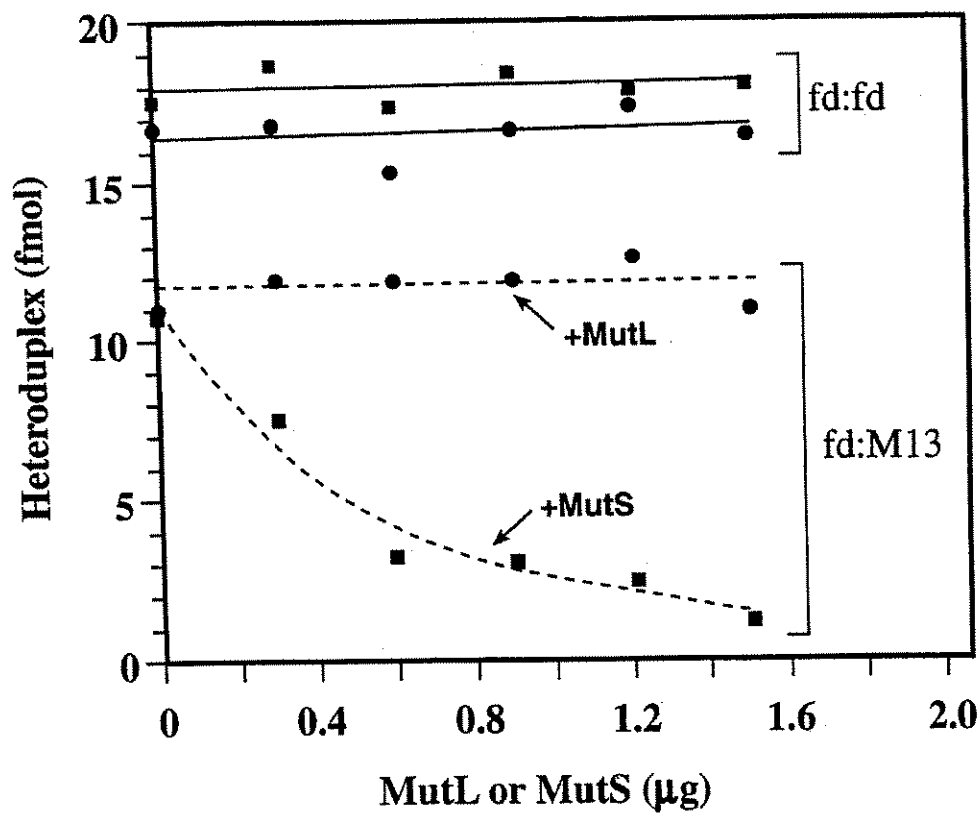


FIG. 7.



U.S. Patent

Sep. 17, 1996

Sheet 6 of 7

5,556,750

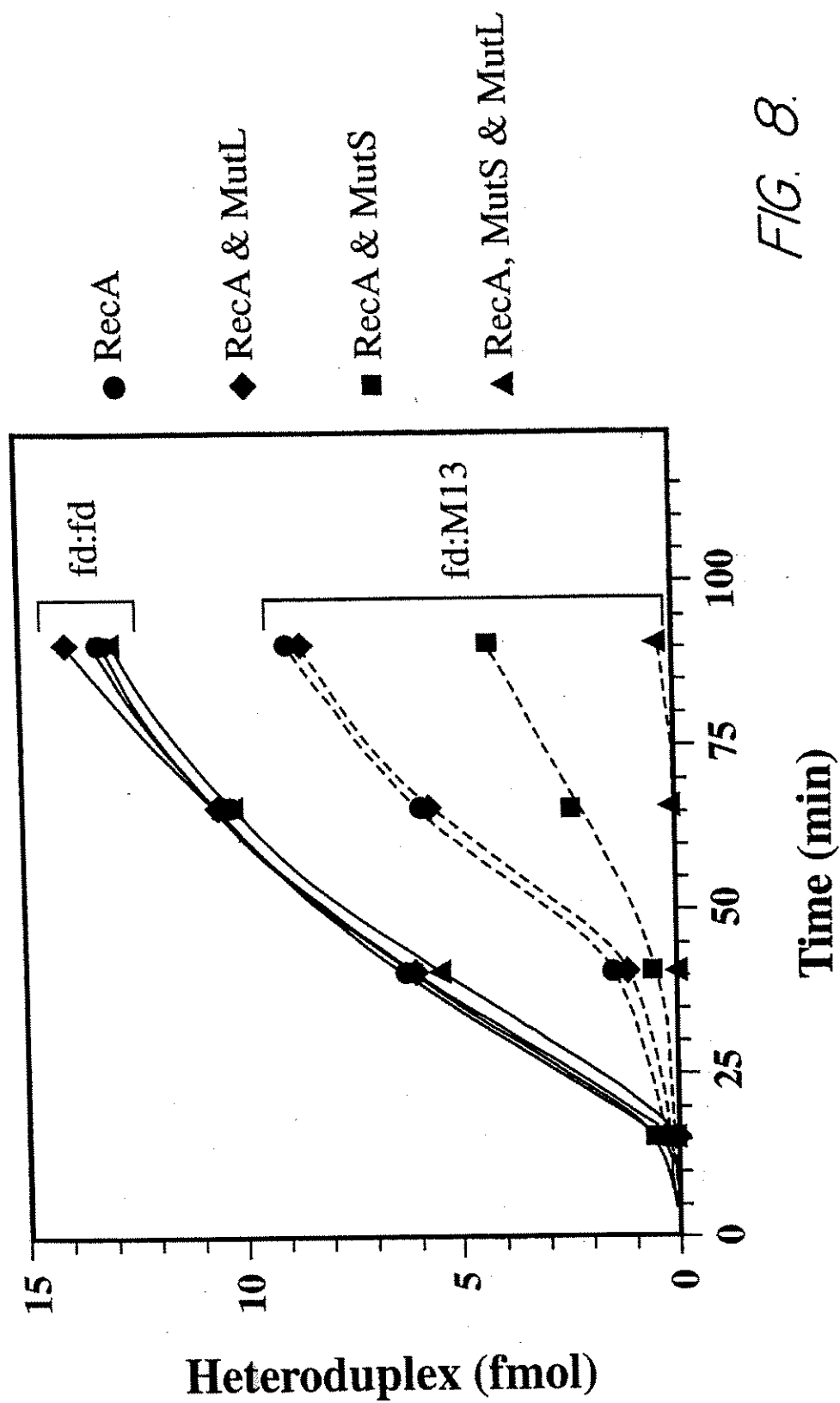


FIG. 8.

U.S. Patent

Sep. 17, 1996

Sheet 7 of 7

5,556,750

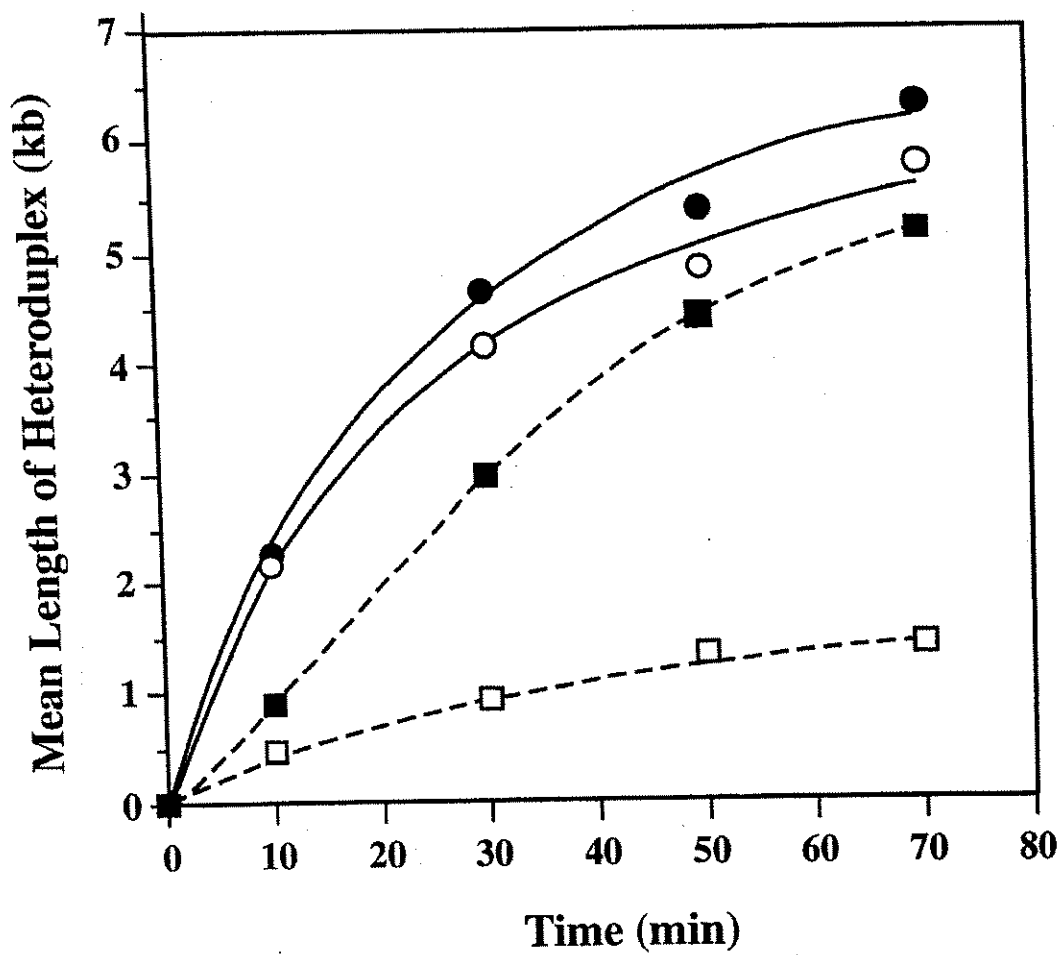


FIG. 9.

5,556,750

1

METHODS AND KITS FOR FRACTIONATING A POPULATION OF DNA MOLECULES BASED ON THE PRESENCE OR ABSENCE OF A BASE-PAIR MISMATCH UTILIZING MISMATCH REPAIR SYSTEMS

DESCRIPTION

This work was supported by the U.S. government, namely Grant No. GM23719. The U.S. government may have rights in this invention.

This application is a continuation-in-part of U.S. application Ser. No. 08/002,529; filed Jan. 11, 1993 and now abandoned; which is a continuation of U.S. application Ser. No. 07/350,983; filed May 12, 1989 and now abandoned.

FIELD OF THE INVENTION

The present invention relates to methods for mapping genetic differences among deoxyribonucleic acid ("DNA") molecules, especially mutations involving a difference in a single base between the base sequences of two homologous DNA molecules.

BACKGROUND OF THE INVENTION

The following is a discussion of relevant art, none of which is admitted to be prior art to the appended claims.

Mapping of genetic differences between individuals is of growing importance for both forensic and medical applications. For example, DNA "fingerprinting" methods are being applied for identification of perpetrators of crimes where even small amounts of blood or sperm are available for analysis. Biological parents can also be identified by comparing DNAs of a child and a suspected parent using such means. Further, a number of inherited pathological conditions may be diagnosed before onset of symptoms, even in utero, using methods for structural analyses of DNA. Finally, it is notable that a major international effort to physically map and, ultimately, to determine the sequence of bases in the DNA encoding the entire human genome is now underway and gaining momentum in both institutional and commercial settings.

DNA molecules are linear polymers of subunits called nucleotides. Each nucleotide comprises a common cyclic sugar molecule, which in DNA is linked by phosphate groups on opposite sides to the sugars of adjoining nucleotides, and one of several cyclic substituents called bases. The four bases commonly found in DNAs from natural sources are adenine, guanine, cytosine and thymine, hereinafter referred to as A, G, C and T, respectively. The linear sequence of these bases in the DNA of an individual encodes the genetic information that determines the heritable characteristics of that individual.

In double-stranded DNA, such as occurs in the chromosomes of all cellular organisms, the two DNA strands are entwined in a precise helical configuration with the bases projecting inward and so aligned as to allow interactions between bases from opposing strands. The two strands are held together in precise alignment mainly by hydrogen bonds which are permitted between bases by a complementarity of structures of specific pairs of bases. This structural complementarity is determined by the chemical natures and locations of substituents on each of the bases. Thus, in double-stranded DNA, normally each A on one strand pairs with a T from the opposing strand, and, likewise, each G with an opposing C.

2

When a cell undergoes reproduction, its DNA molecules are replicated and precise copies are passed on to its descendants. The linear base sequence of a DNA molecule is maintained in the progeny during replication in the first instance by the complementary base pairings which allow each strand of the DNA duplex to serve as a template to align free nucleotides with its polymerized nucleotides. The complementary nucleotides so aligned are biochemically polymerized into a new DNA strand with a base sequence that is entirely complementary to that of the template strand.

Occasionally, an incorrect base pairing does occur during replication, which, after further replication of the new strand, results in a double-stranded DNA offspring with a sequence containing a heritable single base difference from that of the parent DNA molecule. Such heritable changes are called genetic mutations, or more particularly in the present case, "single base pair" or "point" mutations. The consequences of a point mutation may range from negligible to lethal, depending on the location and effect of the sequence change in relation to the genetic information encoded by the DNA.

The bases A and G are of a class of compounds called purines, while T and C are pyrimidines. Whereas the normal base pairings in DNA (A with T, G with C) involve one purine and one pyrimidine, the most common single base mutations involve substitution of one purine or pyrimidine for the other (e.g., A for G or C for T or vice versa), a type of mutation referred to as a "transition". Mutations in which a purine is substituted for a pyrimidine, or vice versa, are less frequently occurring and are called "transversions". Still less common are point mutations comprising the addition or loss of a small number (1, 2 or 3) of nucleotides arising in one strand of a DNA duplex at some stage of the replication process. Such mutations are called small "insertions" or "deletions", respectively, and are also known as "frameshift" mutations in the case of insertion/deletion of one of two nucleotides, due to their effects on translation of the genetic code into proteins. Mutations involving larger sequence rearrangement also do occur and can be important in medical genetics, but their occurrences are relatively rare compared to the classes summarized above.

Mapping of genetic mutations involves both the detection of sequence differences between DNA molecules comprising substantially identical (i.e., homologous) base sequences, and also the physical localization of those differences within some subset of the sequences in the molecules being compared. In principle, it is possible to both detect and localize limited genetic differences, including point mutations within genetic sequences of two individuals, by directly comparing the sequences of the bases in their DNA molecules.

Other methods for detecting differences between DNA sequences have been developed. For example, some pairs of single-stranded DNA fragments with sequences differing in a single base may be distinguished by their different migration rates in electric fields, as in denaturing gradient gel electrophoresis.

DNA restriction systems found in bacteria for example, comprise proteins which generally recognize specific sequences in double-stranded DNA composed of 4 to 6 or more base pairs. In the absence of certain modifications (e.g., a covalently attached methyl group) at definite positions within the restriction recognition sequence, endonuclease components of the restriction system will cleave both strands of a DNA molecule at specific sites within or near the recognition sequence. Such short recognition sequences

5,556,750

3

occur by chance in all natural DNA sequences, once in every few hundred or thousand base pairs, depending on the recognition sequence length. Thus, digestion of a DNA molecule with various restriction endonucleases, followed by analyses of the sizes of the resulting fragments (e.g., by gel electrophoresis), may be used to generate a physical map ("fingerprint") of the locations in a DNA molecule of selected short sequences.

Comparisons of such restriction maps of two homologous DNA sequences can reveal differences within those specific sequences that are recognized by those restriction enzymes used in the available maps. Restriction map comparisons may localize any detectable differences within limits defined ultimately by the resolving power of DNA fragment size determination, essentially within about the length of the restriction recognition sequence under certain conditions of gel electrophoresis.

In practice, selected heritable differences in restriction fragment lengths (i.e., restriction fragment length polymorphisms, "RFLP"s) have been extremely useful, for instance, for generating physical maps of the human genome on which genetic defects may be located with a relatively low precision of hundreds or, sometimes, tens of thousands of base pairs. Typically, RFLPs are detected in human DNA isolated from small tissue or blood samples by using radioactively labeled DNA fragments complementary to the genes of interest. These "probes" are allowed to form DNA duplexes with restriction fragments of the human DNA after separation by electrophoresis, and the resulting radioactive duplex fragments are visualized by exposure to photographic (e.g., X-ray sensitive) film, thereby allowing selective detection of only the relevant gene sequences amid the myriad of others in the genomic DNA.

When the search for DNA sequence differences can be confined to specific regions of known sequence, the recently developed "polymerase chain reaction" ("PCR") technology can be used. Briefly, this method utilizes short DNA fragments complementary to sequences on either side of the location to be analyzed to serve as points of initiation for DNA synthesis (i.e., "primers") by purified DNA polymerase. The resulting cyclic process of DNA synthesis results in massive biochemical amplification of the sequences selected for analysis, which then may be easily detected and, if desired, further analyzed, for example, by restriction mapping or direct DNA sequencing methods. In this way, selected regions of a human gene comprising a few kbp may be amplified and examined for sequence variations.

Another known method for detecting and localizing single base differences within homologous DNA molecules involves the use of a radiolabeled RNA fragment with base sequence complementary to one of the DNAs and a nuclease that recognizes and cleaves single-stranded RNA. The structure of RNA is highly similar to DNA, except for a different sugar and the presence of uracil (U) in place of T; hence, RNA and DNA strands with complementary sequences can form helical duplexes ("DNA:RNA hybrids") similar to double-stranded DNA, with base pairing between A's and U's instead of A's and T's. It is known that the enzyme ribonuclease A ("RNase A") can recognize some single pairs of mismatched bases (i.e., "base mispairs") in DNA:RNA hybrids and can cleave the RNA strand at the mispair site. Analysis of the sizes of the products resulting from RNase A digestion allows localization of single base mismatches, potentially to the precise sequence position, within lengths of homologous sequences determined by the limits of resolution of the RNA sizing analysis (Myers, R. M. et al., 1985, Science, 230, 1242-1246). RNA sizing is performed in this

4

method by standard gel electrophoresis procedures used in DNA sequencing.

S1 nuclease, an endonuclease specific for single-stranded nucleic acids, can recognize and cleave limited regions of mismatched base pairs in DNA:DNA or DNA:RNA duplexes. A mismatch of at least about 4 consecutive base pairs actually is generally required for recognition and cleavage of a duplex by S1 nuclease.

Ford et al., (U.S. Pat. No. 4,794,075) disclose a chemical modification procedure to detect and localize mispaired guanines and thymidines and to fractionate a pool of hybrid DNA from two samples obtained from related individuals. Carbodiimide is used to specifically derivatize unpaired G's and T's, which remain covalently associated with the DNA helix.

The present invention concerns use of proteins that function biologically to recognize mismatched base pairs in double-stranded DNA (and, therefore, are called "mispair recognition proteins") and their application in defined systems for detecting and mapping point mutations in DNAs. Accordingly, it is an object of the present invention to provide methods for using such mispair recognition proteins, alone or in combination with other proteins, for detecting and localizing base pair mismatches in duplex DNA molecules, particularly those DNAs comprising several kbp, and manipulating molecules containing such mismatches. Additionally, it is an object of this invention to develop modified forms of mispair recognition proteins to further simplify methods for identifying specific bases which differ between DNAs. The following is a brief outline of the art regarding mispair recognition proteins and systems, none of which is admitted to be prior art to the present invention.

Enzymatic systems capable of recognition and correction of base pairing errors within the DNA helix have been demonstrated in bacteria, fungi and mammalian cells, but the mechanisms and functions of mismatch correction are best understood in *Escherichia coli*. One of the several mismatch repair systems that have been identified in *E. coli* is the methyl-directed pathway for repair of DNA biosynthetic errors. The fidelity of DNA replication in *E. coli* is enhanced 100-1000 fold by this post-replication mismatch correction system. This system processes base pairing errors within the helix in a strand-specific manner by exploiting patterns of DNA methylation. Since DNA methylation is a post-synthetic modification, newly synthesized strands temporarily exist in an unmethylated state, with the transient absence of adenine methylation on GATC sequences directing mismatch correction to new DNA strands within the hemimethylated duplexes.

In vivo analyses in *E. coli* have shown that selected examples of each of the different mismatches are subject to correction with different efficiencies. G-T, A-C, G-G and A-A mismatches are typically subject to efficient repair. A-G, C-T, T-T and C-C are weaker substrates, but well repaired exceptions exist within this class. The sequence environment of a mismatched base pair may be an important factor in determining the efficiency of repair in vivo. The mismatch correction system is also capable in vivo of correcting differences between duplexed strands involving a single base insertion or deletion. Further, genetic analyses have demonstrated that the mismatch correction process requires intact genes for several proteins, including the products of the mutH, mutL and mutS genes, as well as DNA helicase II and single-stranded DNA binding protein (SSB). The following are further examples of art discussing this subject matter.

5,556,750

5

Lu et al., 80 *Proc. Natl. Acad. Sci. USA* 4639, 1983 disclose the use of a soluble *E. coli* system to support mismatch correction in vitro.

Pans et al., 163 *J. Bact.* 1007, 1985 disclose cloning of the mutS and mutL genes of *Salmonella typhimurium*.

The specific components of the *E. coli* mispair correction system have been isolated and the biochemical functions determined. Preparation of MutS protein substantially free of other proteins has been reported (Su and Modrich, 1986, *Proc. Nat. Acad. Sci. U.S.A.*, 84, 5057-5061, which is hereby incorporated herein by reference). The isolated MutS protein was shown to recognize four of the eight possible mismatched base pairs (specifically, G-T, A-C, A-G and C-T mispairs).

Suet al., 263 *J. Biol. Chem.* 6829, 1988 disclose that the mutS gene product binds to each of the eight base pair mismatches and does so with differential efficiency.

Jiricny et al., 16 *Nucleic Acids Research* 7843, 1988 disclose binding of the mutS gene product of *E. coli* to synthetic DNA duplexes containing mismatches to correlate recognition of mispairs and efficiency of correction in vivo. Nitrocellulose filter binding assays and band-shift assays were utilized.

Welsh et al., 262 *J. Biol. Chem.* 15624, 1987 purified the product of the MutH gene to near homogeneity and demonstrated the MutH gene product to be responsible for d(GATC) site recognition and to possess a latent endonuclease that incises the unmethylated strand of hemimethylated DNA 5' to the G of d(GATC) sequences.

Au et al., 267 *J. Biol. Chem.* 12142, 1992 indicate that activation of the MutH endonuclease requires MutS, MutL and ATP.

Grilley et al. 264 *J. Biol. Chem.* 1000, 1989 purified the *E. coli* mutL gene product to near homogeneity and indicate that the mutL gene product interacts with MutS heteroduplex DNA complex.

Lahue et al., 245 *Science* 160, 1989 delineate the components of the *E. coli* methyl-directed mismatch repair system that function in vitro to correct seven of the eight possible base pair mismatches. Such a reconstituted system consists of MutH, MutL, and MutS proteins, DNA helicase II, single-strand DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, DNA ligase, ATP, and the four deoxyribonucleoside triphosphates.

Su et al., 31 *Genome* 104, 1989 indicate that under conditions of restricted DNA synthesis, or limiting concentration of dNTPs, or by supplementing a reaction with a ddNTP, there is the formation of excision tracts consisting of single-stranded gaps in the region of the molecule containing a mismatch and a d(GATC) site.

Grilley et al. 268 *J. Biol. Chem.* 11830, 1993, indicate that excision tracts span the shorter distance between a mismatch and the d(GATC) site, indicating a bidirectional capacity of the methyl-directed system.

Holmes et al., 87 *Proc. Natl. Acad. Sci. USA*, 5837, 1990, disclose nuclear extracts derived from Hela and *Drosophila melanogaster* K_c cell lines to support strand mismatch correction in vitro.

Cooper et al., 268 *J. Biol. Chem.*, 11823, 1993, describe a role for RecJ and Exonuclease VII as a 5' to 3' exonuclease in a mismatch repair reaction. In reconstituted systems such a 5' to 3' exonuclease function had been provided by certain preparations of DNA polymerase III holoenzyme.

Au et al., 86 *Proc. Natl. Acad. Sci. USA* 8877, 1989 describe purification of the MutY gene product of *E. coli* to

6

near homogeneity, and state that the MutY protein is a DNA glycosylase that hydrolyzes the glycosyl bond linking a mispaired adenine (G-A) to deoxyribose. The MutY protein, an apurinic endonuclease, DNA polymerase I, and DNA ligase were shown to reconstitute G-A to G-C mismatch correction in vitro.

A role for the *E. coli* mismatch repair system in controlling recombination between related but non allelic sequences has been indicated (Feinstein and Low, 113 *Genetics* 13, 1986; Rayssiguier, 342 *Nature* 396, 1989; Shen, 218 *Mol. Gen. Genetics* 358, 1989; Petit, 129 *Genetics* 327, 1991). The frequency of crossovers between sequences which differ by a few percent or more at the base pair level are rare. In bacterial mutants deficient in methyl-directed mismatch repair, the frequency of such events increases dramatically. The largest increases are observed in MutS and MutL deficient strains. (Rayssiguier, supra; and Petit, supra.)

Nelson et al., 4 *Nature Genetics* 11, 1993, disclose a genomic mismatch (GMS) method for genetic linkage analysis. The method allows DNA fragments from regions of identity-by-descent between two relatives to be isolated based on their ability to form mismatch-free hybrid molecules.

The method consists of digesting DNA from the two sources with a restriction endonuclease that produces protruding 3' ends. The protruding 3' ends provide some protection from exonuclease III, which is used in later steps. The two sources are distinguished by methylating the DNA from only one source. Molecules from both sources are denatured and reannealed, resulting in the formation of four types of duplex molecules: homohybrids formed from strands derived from the same source and heterohybrids consisting of DNA strands from different sources. Heterohybrids can either be mismatch free or contain base-pair mismatches, depending on the extent of identity of homologous regions.

Homohybrids are distinguished from heterohybrids by use of restriction endonucleases that cleave at fully methylated or unmethylated GATC sites. Homohybrids are cleaved to smaller duplex molecules, while heterohybrid are resistant to cleavage. Heterohybrids containing a mismatch(es) are distinguished from mismatch free molecules by use of the *E. coli* methyl-directed mismatch repair system. The combination of three proteins of the methyl-directed mismatch repair system MutH, MutL, and MutS along with ATP introduce a single-strand nick on the unmethylated strand at GATC sites in duplexes that contain a mismatch. Heterohybrids that do not contain a mismatch are not nicked. All molecules are then subject to digestion by Exonuclease III (Exo III), which can initiate digestion at a nick, a blunt end or a 5' overhang, to produce single-stranded gaps. Only mismatch free heterohybrids are not subject to attack by Exo III, all other molecules have single-stranded gaps introduced by the enzyme. Molecules with single-stranded regions are removed by absorption to benzoylated naphthoylated DEAE cellulose. The remaining molecules consist of mismatch-free heterohybrids which may represent regions of identity by descent.

SUMMARY OF THE INVENTION

Applicant has determined that a single DNA base mispair recognition protein can form specific complexes with any of the eight possible mismatched base pairs embedded in an otherwise homologous DNA duplex. It has also been revealed that another mispair recognition protein can rec-

5,556,750

7

ognize primarily one specific base pair mismatch, A-G, and in so doing, it chemically modifies a nucleotide at the site of the mispair. In addition, defined in vitro systems have been established for carrying out methyl-directed mismatch repair processes. Accordingly, the present invention features the use of such mispair recognition proteins and related correction system components to detect and to localize point mutations in DNAs. In addition the invention concerns methods for the analysis and manipulation of populations of DNA duplex molecules potentially containing base pair mismatches through the use of all or part of defined mismatch repair systems.

The invention utilizes five basic methods for heteroduplex mapping analysis, and manipulation: (i) binding of a mismatch recognition protein, e.g., MutS to DNA molecules containing one or more mispairs; (ii) cleavage of a heteroduplex in the vicinity of a mismatch by a modified form of a mismatch recognition protein; (iii) mismatch-provoked cleavage at one or more GATC sites via a mismatch repair system dependent reaction, e.g., MthLS; (iv) formation of a mismatch-provoked gap in heteroduplex DNA via reactions of a mismatch repair system and (v) labelling of mismatch-containing nucleotides with a nucleotide analog, e.g., a biotinylated nucleotide, using a complete mismatch repair system.

For clarity in the following discussion, it should be noted that certain distinctions exist related to the fact that some proteins that recognize DNA base mispairs are merely DNA binding proteins, while others modify the DNA as a consequence of mispair recognition. Notwithstanding the fact that in the latter situation the protein modifying the DNA may be associated with the DNA only transiently, hereinafter, whether a mispair recognition protein is capable of DNA binding only or also of modifying DNA, whenever it is said that a protein recognizes a DNA mispair, this is equivalent to saying that it "forms specific complexes with" or "binds specifically to" that DNA mispair in double-stranded DNA. In the absence of express reference to modification of DNA, reference to DNA mispair recognition does not imply consequent modification of the DNA. Further, the phrase "directs modification of DNA" includes both cases wherein a DNA mispair recognition protein has an inherent DNA modification function (e.g., a glycosylase) and cases wherein the mispair recognition protein merely forms specific complexes with mispairs, which complexes are then recognized by other proteins that modify the DNA in the vicinity of the complex.

Accordingly, the present invention features a method for detecting base pair mismatches in a DNA duplex by utilizing a mismatch recognition protein that forms specific complexes with mispairs, and detecting the resulting DNA:protein complexes by a suitable analytical method.

In addition to methods designed merely to detect base pair mismatches, this invention includes methods for both detecting and localizing base pair mismatches by utilizing components of mismatch repair system.

The present invention also features mispair recognition proteins which have been altered to provide an inherent means for modifying at least one strand of the DNA duplex in the vicinity of the bound mispair recognition protein.

The present invention also concerns systems utilizing an A-G specific mispair recognition protein, for example, the *E. coli* DNA mispair recognition protein that recognizes only A-G mispairs without any apparent requirement for hemimethylation. This protein, the product of the mutY gene, is a glycosylase which specifically removes the

8

adenine from an A-G mispair in a DNA duplex. Accordingly, this MutY protein is useful for the specific detection of A-G mispairs according to the practice of the present invention.

The invention also includes the combined use of components of a mismatch repair system along with a recombinase protein. The recombinase protein functions to catalyze the formation of duplex molecules starting with single-stranded molecules obtained from different sources, by a renaturation reaction. Such a recombinase protein is also capable of catalyzing a strand transfer reaction between a single-stranded molecule from one source and double-stranded molecules obtained from a different source. In the presence of a base pair mismatch, formation of duplex regions catalyzed by such a recombinase protein is inhibited by components of a mismatch repair system, e.g., *E. coli* MutS and MutL, proteins. Modulation of recombinase activity by components of a mismatch repair system may involve inhibition of branch migration through regions that generate mismatched base pairs. The combination of a DNA mismatch repair system and a recombinase system provides a very sensitive selection step allowing for the removal of molecules containing a base pair mismatch from a population of newly formed heteroduplex molecules. This procedure provides a selection scheme that can be utilized independent of or in conjunction with the actual mismatch repair reaction.

The invention also features two improvements on the genomic mismatch scanning technique (GMS) of Nelson et al. 4 *Nature Genetics* 11, 1993, used to map regions of genetic identity between populations of DNA molecules.

One improvement provided by the invention features an additional selection step, as described above, for determining genetic variation. The genomic mismatch scanning (GMS) method includes one selection step which is carried out after hybrid formation. The present invention includes an additional step that occurs during hybrid formation, through the use of a protein with recombinase activity along with components of a mismatch repair system. The increase in sensitivity for screening for genetic variation provided by the additional selection step makes possible the use of the GMS technique with larger genomes, e.g., man.

A second improvement provided by the invention features the replacement or modification of the exonuclease III digestion step employed in the GMS method. In the GMS procedure exonuclease III is used to degrade all DNA molecules, except mismatch-free heterohybrids, to molecules containing single-stranded regions, which are subsequently removed. Heterohybrids are duplex molecules which are formed in the method from two molecules which were previously base paired with other molecules (i.e., from different sources). In the instant invention this step is replaced by a procedure that employs all or some of the components of a mismatch repair system. Exo III is a 3' to 5' exonuclease specific for double-stranded DNA, which preferably initiates at blunt or 5' protruding ends. In the GMS procedure DNA molecules are digested with restriction enzymes that produce protruding 3' ends. Although molecules containing protruding 3' ends are not preferred substrates for Exo III, such molecules can be subject to limited attack by the enzyme. Thus, even mismatch-free heterohybrids will be degraded to some extent by Exo III, and will be erroneously removed from the final population of molecules representing those of identity-by-descent. The invention employs components of a mismatch repair system along with dideoxy or biotinylated nucleotide, to avoid the use of Exo III and the potential loss of heterohybrids

5,556,750

9

molecules that are mismatch-free. Homohybrids are digested in the presence of helicase II by *exoVI* RecJ and *exo I*, e.g., natural exonucleases involved in the mismatch repair reaction. The invention also features a modification of the step utilizing *Exo III*, consisting of ligation of duplex DNA molecules at dilute concentrations so as to form closed circular monomer molecules, thus removing any 3' ends which may be subject to degradation by *Exo III*.

The invention includes the use of a mismatch repair system to detect and remove or correct base pair mismatches in a population produced by the process of enzymatic amplification of nucleic acid molecules. DNA polymerase errors that occur during a cycle of enzymatic amplification can result in the presence of mismatched base pair(s) in the population of product molecules. If such errors are perpetuated in subsequent cycles they can impair the value of the final amplified product. The fidelity of the amplification method can be enhanced by including one or more components of a mismatch repair system to either correct the mismatch base pair(s) or to eliminate from the amplified population, molecules that contain mismatch base pair(s). Elimination of molecules containing a base pair mismatch can be accomplished by binding to a protein, such as MutS, or by introduction of a nick in one strand of the duplex so that a full sized product will not be produced in a subsequent round of amplification.

The invention also features methods to remove molecules containing a base pair mismatch through the binding of the mismatch to the components of the mismatch repair system or by the binding of a complex of a mismatch and components of a mismatch repair system to other cellular proteins. Another aspect of the invention for removal of molecules containing a mismatch is through the incorporation of biotin into such a molecule and subsequent removal by binding to avidin.

Another aspect of the invention features use of a mismatch repair system which has a defined 5' to 3' exonuclease function, that is provided by the exonuclease VII or RecJ exonuclease. In other systems a 5' to 3' exonuclease function is provided by exonuclease VII which is present in many preparations of the DNA polymerase III holoenzyme.

The invention also includes kits having components necessary to carry out the methods of the invention.

The mismatch repair systems of the instant invention, e.g., *E. coli*, offer specific and efficient procedures for detection and localization of mismatches and manipulation of DNA containing mismatches that is a reflection of their biological function. All eight possible base pair mismatches are recognized and seven of the eight mismatches are processed and corrected by the system. Although C-C mismatches are not a substrate for repair, MutS does bind weakly to this mismatch permitting its detection. In contrast to the electrophoretic migration procedure, the RNase method, or chemical modification procedures, the system does not depend on the destabilization of the DNA helix for detection of mismatches or binding to mismatches. The system features exquisite specificity, and is not subject to non-specific interactions with bases at the ends of linear DNA fragments or non-specific interactions at non-mismatch sites in long molecules.

The detection of fragments containing a mispair is limited only by the intrinsic specificity of the system, for example, detection of better than one G-T mispair per 300 kilobases. Mismatches have been routinely detected with a 6,400 base pair substrate and the system should be applicable to molecules as large as 40-50 kb. This allows for detection of

10

possible single base differences between long DNA sequences, for example, between a complete gene from one individual and the entire genome of another. The invention also enables the localization of any possible single base difference within the sequences of homologous regions of long DNA molecules such as those encoding one or more complete genes and comprising several kbp of DNA.

Several of the methods of the invention result in the covalent alteration of the phosphodiester backbone of DNA molecules. This covalent alteration facilitates analysis of the product DNA molecules especially by electrophoretic methods.

Other features and advantages of the invention will be apparent from the following description of the preferred embodiments thereof, and from the claims.

BRIEF DESCRIPTION OF THE FIGURES

FIG. 1. Heteroduplex substrate for in vitro mismatch correction. The substrate used in some examples is a 6440-bp, covalently closed, circular heteroduplex that is derived from bacteriophage ϕ 1 and contains a single base-base mismatch located within overlapping recognition sites for two restriction endonucleases at position 5632. In the example shown a G-T mismatch resides within overlapping sequences recognized by Hind III and Xho I endonucleases. Although the presence of the mispair renders this site resistant to cleavage by either endonuclease, repair occurring on the complementary (c) DNA strand yields an A-T base pair and generates a Hind III-sensitive site, while correction on the viral (v) strand results in a G-C pair and Xho I-sensitivity. The heteroduplexes also contain a single d(GATC) sequence 1024 base pairs from the mismatch (shorter path) at position 216. The state of strand methylation at this site can be controlled, thus permitting evaluation of the effect of DNA methylation on the strand specificity of correction.

FIG. 2. Requirement for DNA ligase in mismatch correction. Hemimethylated G-T heteroduplex DNA (FIG. 1, 0.6 μ g, d(GATC) methylation on the complementary DNA strand) was subjected to mismatch repair under reconstituted conditions in a 60 μ l reaction (Table 3, closed circular heteroduplex), or in 20 μ l reactions (0.2 μ g of DNA) lacking MutS protein or ligase, or lacking both activities. A portion of each reaction (0.1 μ g of DNA) was treated with EDTA (10 mM final concentration) and subjected to agarose gel electrophoresis in the presence of ethidium bromide (1.5 μ g/ml; top panel, lanes 1-4). Positions are indicated for the unreacted, supercoiled substrate (SC), open circles containing a strand break (OC) and covalently closed, relaxed circular molecules (RC). A second sample of each reaction containing 0.1 μ g of DNA was hydrolyzed with Xho I and Cla I endonucleases (FIG. 1) to score G-T to G-C mismatch correction and subjected to electrophoresis in parallel with the samples described above (bottom panel, lanes 5-8). The remainder of the complete reaction (0.4 μ g DNA, corresponding to the sample analyzed in lane 1) was made 10 mM in EDTA, and subjected to electrophoresis as described above. A gel slice containing closed circular, relaxed molecules was excised and the DNA eluted. This sample was cleaved with Xho I and Cla I and the products analyzed by electrophoresis (lane 9).

FIG. 3. Methyl-direction of mismatch correction in the purified system. Repair reactions with the G-T heteroduplex (FIG. 1) were performed as described in Table 3 (closed circular heteroduplex) except that reaction volumes were 20

5,556,750

11

μl (0.2 μg of DNA) and the incubation period was 60 minutes. The reactions were heated to 55° for 10 minutes and each was divided into two portions to test strand specificity of repair. G-T to A-T mismatch correction, in which repair occurred on the complementary (c) DNA strand, was scored by cleavage with Hind III and Cla I endonucleases, while hydrolysis with Xho I and Cla I were used to detect G-T to G-C repair occurring on the Viral (v) strand. Apart from the samples shown in the left two lanes, all heteroduplexes were identical except for the state of methylation of the single d(GATC) sequence at position 216 (FIG. 1). The state of modification of the two DNA strands at this site is indicated by + and - notation. The G-T heteroduplex used in the experiment shown in the left two lanes (designated 0/0) contains the sequence d(GATT) instead of d(GATC) at position 216, but is otherwise identical in sequence to the other substrates.

FIG. 4. Strand-specific repair of heteroduplexes containing a single strand scission in the absence of MutH protein. Hemimethylated G-T heteroduplex DNAs (FIG. 1, 5 μg) bearing d(GATC) modification on the viral or complementary strand were subjected to site-specific cleavage with near homogeneous MutH protein. Because the MutH-associated endonuclease is extremely weak in the absence of other mismatch repair proteins, cleavage at d(GATC) sites by the purified protein requires a MutH concentration 80 times that used in reconstitution reactions. After removal of MutH by phenol extraction, DNA was ethanol precipitated, collected by centrifugation, dried under vacuum, and resuspended in 10 mM Tris-HCl (pH 7.6), 1 mM EDTA. Mismatch correction of MutH-incised and covalently closed, control heteroduplexes was performed as described in the legend to Table 2 except that ligase and NAD⁺ were omitted. Outside and inside strands of the heteroduplexes depicted here correspond to complementary and viral strands respectively. Values in parentheses indicate repair occurring on the methylated, continuous DNA strand. The absence of MutH protein in preparations of incised heteroduplexes was confirmed in two ways. Preparations of incised molecules were subject to closure by DNA ligase (>80%) demonstrating that MutH protein does not remain tightly bound to incised d(GATC) sites. Further, control experiments in which each MutH-incised heteroduplex was mixed with a closed circular substrate showed that only the open circular form was repaired if MutH protein was omitted from the reaction whereas both substrates were corrected if MutH protein was present (data not shown).

FIG. 5. Requirements for MutH protein and a d(GATC) sequence for correction in the presence of DNA ligase. Hemimethylated G-T heteroduplexes incised on the unmethylated strand at the d(GATC) sequence were prepared as described above in FIG. 4. A G-T heteroduplex devoid of d(GATC) sites (FIG. 4) and containing a single-strand break within the complementary DNA strand at the Hinc II site (position 1) was constructed as described previously (Lahue et al. supra). Mismatch correction assays were performed as described in Table 3, with ligase (20 ng in the presence of 25 μM NAD⁺) and MutH protein (0.26 ng) present as indicated. Table entries correspond to correction occurring on the incised DNA strand, with parenthetical values indicating the extent of repair on the continuous strand. Although not shown, repair of the nicked molecule lacking a d(GATC) sequence (first entry of column 3) was reduced more than an order of magnitude upon omission of MutL, MutS, SSB or DNA polymerase III holoenzyme.

FIG. 6 is a diagrammatic representation of the model system used to evaluate MutS and MutL effects on RecA catalyzed strand transfer.

12

FIG. 7 depicts the effects of MutS and MutL on RecA-catalyzed strand transfer between homologous and quasi-homologous DNA sequences. Solid lines indicate fd-fd strand transfer, while dashed lines correspond to fd-M13 strand transfer. Strand transfer was evaluated in the presence of MutL (solid circles) or MutS (solid squares).

FIG. 8 depicts The MutL potentiation of MutS block to strand transfer in response to mismatched base pairs. Solid lines: fd-fd strand transfer; dashed lines fd-M13 strand transfer; RecA (solid circle); RecA and MutL (solid diamond); RecA and MutS (solid square); RecA, MutL, and MutS (solid triangle).

FIG. 9 depicts the MutS and MutL block of branch migration through regions that generate mismatched base pairs. Solid lines: M13-M13 strand transfer; dashed line fd-M13 strand transfer. RecA only (solid circle and square); RecA, MutS, and MutL (open circle and square).

DESCRIPTION OF PREFERRED EMBODIMENTS

The invention consists of methods utilizing and kits consisting of components of mismatch repair system to detect, and localize DNA base pair mismatches and manipulate molecules containing such mismatches. The invention also features modified mispair recognition proteins and their utilization in the above-mentioned methods and kits. The invention also includes methods and kits comprising components of a mismatch repair system along with proteins with recombinase activity. The invention also consists of methods to improve the GMS technique to detect regions of homology-by-descent.

Methods for Detecting the Presence and Localization of Mismatched Base Pairs by Complex Formation with a Mismatch Recognition Protein

One embodiment of the invention features a diagnostic method for detecting a base pair mismatch in a DNA duplex. The method comprises the steps of contacting at least one strand of a first DNA molecule with the complementary strand of a second DNA molecule under conditions such that base pairing occurs, contacting a DNA duplex potentially containing a base pair mismatch with a mispair recognition protein under conditions suitable for the protein to form a specific complex only with the DNA duplex having a base pair mismatch, and not with a DNA duplex lacking a base pair mismatch, and detecting the complex as a measure of the presence of a base pair mismatch in the DNA duplex.

By "mismatch" is meant an incorrect pairing between the bases of two nucleotides located on complementary strands of DNA, i.e., bases pairs that are not A:T or G:C.

In the practice of this method, the two DNA's or two DNA samples to be compared may comprise natural or synthetic sequences encoding up to the entire genome of an organism, including man, which can be prepared by well known procedures. Detection of base sequence differences according to this method of this invention does not require cleavage (by a restriction nuclease, for example) of either of the two DNAs, although it is well known in the art that rate of base pair formation between complementary single-stranded DNA fragments is inversely related to their size. This detection method requires that base sequence differences, to be detected as base pair mismatches lie within a region of homology constituting at least about 14 consecutive base pairs of homology between the two DNA molecules, which

5,556,750

13

is about the minimum number of base pairs generally required to form a stable DNA duplex. Either one or both of the strands of the first DNA may be selected for examination, while at least one strand of the second DNA complementary to a selected first DNA strand must be used. The DNA strands, particularly those of the second DNA, advantageously may be radioactively labeled to facilitate direct detection, according to procedures well known in the art.

By "mismatch recognition protein" is meant a protein of a mismatch repair system that specifically recognizes and binds to a base pair mismatch, e.g., coli MutS.

Methods and conditions for contacting the DNA strands of the two DNAs under conditions such that base pairing occurs are also widely known in the art.

In preferred embodiments of this aspect of this invention, the mismatch recognition protein is the product of the mutS gene of *E. coli* or species variations thereof, or portions thereof encoding the recognition domain. The protein recognizes all eight possible base pair mismatches, detection of the DNA:protein complex comprises contacting the complexes with a selectively adsorbent agent under conditions such that the protein:DNA complexes are retained on the agent while DNA not complexed with protein is not retained and measuring the amount of DNA in the retained complexes, the adsorbent agent is a membranous nitrocellulose filter, detection of the DNA:protein complex further includes the step wherein an antibody specific for the base mismatch recognition protein is employed, the base mismatch recognition protein is the product of the mutS gene of *S. typhimurium* the hexA gene of *S. pneumoniae* or the MSH1 and MSH2 genes of yeast, and wherein the step for detecting the DNA:protein complex further includes a step wherein the electrophoretic mobility of the DNA:protein complex is compared to uncomplexed DNA.

The ability of the MutS protein to recognize examples of all eight single base pair mismatches within double-stranded DNA, even including C—C mismatches which do not appear to be corrected in vivo, is demonstrated by the fact that MutS protein protects DNA regions containing each mismatch from hydrolysis by DNase I (i.e., by "Dnase I footprint" analyses), as recently reported (Su, S.-S., et al., 1988, *J. Biol. Chem.*, 263, 6829–6835). The affinity of MutS protein for the different mismatches that have been tested varies considerably. Local sequence environment may also affect the affinity of the MutS protein for any given base mismatch; in other words, for example, the affinity for two specific cases of A—C mismatches, which are surrounded by different sequences, may not be the same. Nevertheless, no examples of base mismatches have been found that are not recognized by isolated MutS protein. Accordingly, this method of the invention detects all mismatched base pairs.

It should be particularly noted that the DNA duplexes which MutS recognizes are not required to contain GATC sequences and, hence, they do not require hemimethylation of A's in GATC sequences, the specific signal for the full process of methyl-directed mismatch correction in vivo; therefore, use of MutS in this method allows recognition of a DNA base mismatch in DNAs lacking such methylation, for instance, DNAs isolated from human tissues.

By "species variation" is meant a protein which appears to be functionally and in part, at least, structurally homologous to the *E. coli* MutS protein. One example of such a protein has also been discovered in a methyl-directed mismatch correction system in *Salmonella typhimurium* bacteria (Pang et al., 1985, *J. Bacteriol.*, 163, 1007–1015). The gene for this protein has been shown to complement *E. coli* strains

14

with mutations inactivating the mutS gene and the amino acid sequence of its product shows homology with that of the *E. coli* MutS protein. Accordingly, this *S. typhimurium* protein is also suitable for the practice of this aspect of the present invention. Other organisms, including man, are known to possess various systems for recognition and repair of DNA mismatches, which, as one skilled in the art would appreciate, comprise mismatch recognition proteins functionally homologous to the MutS protein. Nuclear extracts derived from HeLa and *Drosophila melanogaster* K_c cell lines has been shown to support efficient strand-specific specific mismatch correction in vitro (Holmes et al., 1990, *Proc. Natl. Acad. Sci. USA* 87, 5837–5841, which is incorporated herein by reference), and this reaction has been shown to occur by a mechanism similar to that of the bacterial reaction (Fany and Modrich 268 *J. Biol. Chem.* 11838, 1993). Furthermore, genes encoding proteins that are homologous to bacterial MutS at the amino acid sequence level have been demonstrated in human (Fujii and Shimada 264 *J. Biol. Chem.* 10057, 1989) and yeast (Reenan and Kolodner 132 *Genetics* 963, 1992) and *S. pneumoniae* (Priebe et al., 170 *J. Bacteriol.* 190, 1988). Accordingly, it is believed that such DNA base mismatch recognition proteins may also be suitable for use in the present invention.

By "protein encoding the recognition domain" is meant a region of the mismatch recognition protein which is involved in mismatch recognition and binding. Such a domain comprises less than the complete mismatch recognition protein.

By a "selectively adsorbent agent" is meant any solid substrate to which protein:DNA complexes are retained on the agent while DNA not complexed with protein is not retained, such agents are known to those skilled in the art. Absent radioactive labeling of at least one strand used to form the DNA duplexes, the DNA in complexes on the filter may be detected by any of the usual means in the art for detection of DNA on a solid substrate, including annealing with complementary strands of radioactive DNA.

The nitrocellulose filter method for detecting complexes of MutS protein with base mismatches in DNA has been reported in detail (Jiricny, J. et al., 1988, *Nuc. Acids Res.* 16, 7843–7853, which is hereby incorporated herein by reference). Besides simplicity, a major advantage of this method for detecting the DNA:protein complex over other suitable methods is the practical lack of a limitation on the size of DNA molecules that can be detected in DNA:protein duplexes. Therefore, this embodiment of this method is in principle useful for detecting single base sequence differences between DNA fragments as large as can be practically handled without shearing.

By "electrophoretic mobility" is meant a method of separating the DNA:protein complexes from DNA that does not form such complexes on the basis of migration in a gel medium under the influence of an electric field. DNA:protein complexes are less mobile than naked DNA. Such methods based on electrophoretic mobility are known to those skilled in the art. The DNA in the DNA:protein complexes may be detected by any of the usual standard means for detection of DNA in gel electrophoresis, including staining with dyes or annealing with complementary strands of radioactive DNA. Detecting complexes comprising the MutS base mismatch recognition protein and mismatches in DNA duplexes is also described in the foregoing reference (Jiricny, J. et al., 1988, *Nuc. Acids Res.*, 16, 7843–7853). Under the usual conditions employed in the art for detecting specific DNA:protein complexes by gel electrophoresis, complex formation of a protein with a double-stranded DNA fragment of up to several hundred base pairs is known to produce distinguishable mobility differences.

5,556,750

15

Antibodies specific for a DNA mispair recognition protein can be prepared by standard immunological techniques known to those skilled in the art.

Other suitable analytical methods for detecting the DNA protein complex include immunodetection methods using an antibody specific for the base mispair recognition protein. For example, antibodies specific for the *E. coli* MutS protein have been prepared. Accordingly, one immunodetection method for complexes of MutS protein with DNA comprises the steps of separating the DNA:protein complexes from DNA that does not form such complexes by immunoprecipitation with an antibody specific for MutS protein, and detecting the DNA in the precipitate. According to the practice of this aspect of the invention, quantitative immunoassay methods known in the art may be employed to determine the number of single base mispairs in homologous regions of two DNA molecules, based upon calibration curves that can be established using complexes of a given mispair recognition protein with DNA duplexes having known numbers of mispairs.

Another aspect of the invention features a method for detecting and localizing a base pair mismatch in a DNA duplex. The method includes contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting double-stranded DNA duplexes with a mispair recognition protein under conditions such that the protein forms specific complexes with mispairs, subjecting the duplex molecules to hydrolysis with an exonuclease under conditions such that the complex blocks hydrolysis, and determining the location of the block to hydrolysis by a suitable analytic method.

"Hydrolysis with an exonuclease" is a procedure known to those skilled in the art and utilizes enzymes possessing double-strand specific exonuclease activity, e.g., *E. coli* exonuclease III, RecBCD exonuclease, lambda exonuclease, and T7 gene 6 exonuclease.

By "block to hydrolysis" is meant interference of hydrolysis by the exonuclease. Such protection can result from the mispair recognition protein protecting the DNA to which it is bound.

By "suitable analytical method" is meant any method that allows detection of the block to exonuclease digestion, such analysis of molecules by gel electrophoresis. Such methods are known to those skilled in the art.

Methods for Detecting and Localizing Base Pair Mismatches by Mismatch Repair System Strand Modification Reactions

In addition to methods that detect base sequence differences, this invention provides methods for both detecting and localizing a base pair mismatch in a DNA duplex. One method includes contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting double-stranded DNA duplexes with a mismatch recognition protein under conditions such that the protein forms specific complexes with mispairs and thereby directs modification of at least one strand of the DNA in the resulting DNA:protein complexes in the vicinity of the DNA:protein complex, and determination of the location of the resulting DNA modification by a suitable analytical method.

By "modification" is meant any alteration for which there is a means of detection, for instance a chemical modification

16

including breaking of a chemical bond resulting in, as examples, cleavage between nucleotides of at least one DNA strand or removal of a base from the sugar residue of a nucleotide. Specific means for modifying DNAs in the vicinity of the DNA:protein complex are provided below for several embodiments of this aspect of the invention, together with interpretations of the phrase "in the vicinity of", as appropriate to the practical limitations of the modification approach in each instance.

Suitable analytical methods for determining the location of the modification are known to those skilled in the art. Such a determination involves comparison of the modified DNA molecule with the homologous unmodified DNA molecule.

In preferred embodiments of this aspect of the invention, the mispair recognition protein is the product of the mutS gene of *E. coli* or another functionally homologous protein; the step in which the DNA is modified in the vicinity of the DNA:protein complex further comprises contacting the DNA:MutS protein complex with a defined set or subset of *E. coli* DNA mismatch repair proteins (comprising *E. coli* MutH, MutL, DNA helicase II, single-stranded DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, and exonuclease VII (or RecJ exonuclease), or species variations of these activities), ATP and one or more dideoxynucleoside-5'-triphosphates or in the absence of exogenous deoxyribonucleoside-5'-triphosphate under conditions that produce a discontinuity in one or both strands of the DNA duplex in the vicinity of the mismatch.

DNA used in such an analysis is to be unmethylated or hemimethylated at on the 6-position of the adenine base in GATC sequences. With the exception of DNAs from some bacterial species, the chromosomes of most organisms naturally lack this modification. In those cases where hemimethylation of otherwise GATC unmodified molecules is desired, this can be accomplished by use of *E. coli* Dam methylase as is well known in the art. Symmetrically methylated DNA prepared by use of this enzyme is denatured and subsequently reannealed with single-stranded sequences representing an homologous (or largely so) DNA. If necessary, hemimodified molecules produced by this renaturation procedure can be separated from unmethylated is symmetrically methylated duplexes which can also result from the annealing procedure. As is well known in the art, this can be accomplished by subjecting annealed products to cleavage by DpnI and MboI endonucleases. The former activity cleaves symmetrically methylated duplex DNA at GATC sites while unmodified duplex DNA is subject to double strand cleavage only at unmodified GATC sites by the latter activity. Since hemimodified DNA is resistant to double strand cleavage by both DpnI and MboI, desired hemimethylated products can be separated on the basis of size from the smaller fragments produced by DpnI and MboI cleavage, for example by electrophoretic methods.

By "discontinuity in one or both strands of the DNA duplex" is meant a region which consists of a break in the phosphodiester backbone in one or both strands, or a single-stranded gap in a duplex molecule.

One aspect of this preferred embodiment involves contacting the DNA:MutS protein complex with *E. coli* MutL and MutH proteins (or species variations thereof) in the presence of ATP and an appropriate divalent cation cofactor (e.g., Mg^{2+}) so that mismatch-containing molecules will be subject to incision at one or more GATC sites in the vicinity of the mispair. Such incision events can be monitored by a suitable analytic method for size detection such as electrophoresis under denaturing condition.

5,556,750

17

A second aspect of this preferred embodiment involves contacting the DNA:MutS complex with a defined *E. coli* mismatch correction system consisting of *E. coli* MutH, MutL, DNA helicase II, single-stranded DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, and exonuclease VII (or RecJ exonuclease), or species variants of these activities, ATP in the absence of exogenous deoxyribonucleoside-5'-triphosphates or in the presence of one or more dideoxynucleoside-5'-triphosphates such that single-stranded gaps are produced in the vicinity of the complexed protein; the method for determining the location of the single-stranded gaps with the DNA duplex further includes analysis of electrophoretic mobility of treated samples under denaturing conditions of the steps of cleaving the DNA with a single-stranded specific endonuclease, and comparing the electrophoretic mobilities of the cleaved fragments with unmodified DNA fragments under non-denaturing conditions; the step for modifying the DNA duplex in the vicinity of the complexed protein comprises contacting the complexes with proteins of a mismatch repair system, ATP and a divalent cation under conditions such that an endonucleolytic incision is introduced at one or more GATC sequences in the duplex molecule.

An example of a complete defined mismatch correction system comprises the following purified components: *E. coli* MutH, MutL, and MutS proteins, DNA helicase II, single-stranded DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, DNA ligase, ATP, and the four deoxynucleoside-5'-triphosphates. This set of proteins can process seven of the eight base-base mismatches in a strand-specific reaction that is directed by the state of methylation of a single GATC sequence located 1 kilobase from the mispair. This defined system is described further in Example 1, below. The 5' to 3' exonuclease function can either be supplied by either DNA polymerase III holoenzyme preparations that contain this activity or as a separate defined component consisting of exonuclease VII or RecJ exonuclease. It should be noted that the lack of ability to repair C—C base mispairs in this embodiment of this aspect of the present invention is not a major limitation of the method for detecting all possible base sequence differences between any two naturally occurring DNA sequences because mutations that would give rise to a C—C mispair upon hybridization would also give rise to a G—G mismatch when the complementary strands are hybridized.

For the purpose of generating single-stranded gaps in the vicinity of the DNA:MutS protein complexes, DNA duplexes containing mispaired base pairs are contacted with the defined mismatch correction system under the standard conditions described in Example 1, Table 3 (Complete reaction), except for the following differences: (i) exogenous dNTPs are omitted; or (ii) 2', 3'-dideoxynucleoside-5'-triphosphates (ddNTPs) at suitable concentrations (10 to 100 μ M) are substituted for dNTPs; or (iii) reactions containing dNTPs are supplemented with ddNTPs at a suitable concentration to yield a chain termination frequency sufficient to inhibit repair of single-strand gaps. In cases (i)–(iii) DNA ligase may be omitted from the reaction. In cases (ii) and (iii) all four ddNTPs may be present; however, it is expected that the presence of one, two, or three ddNTPs will prove sufficient to stabilize single strand gaps via chain termination events. While it is expected that most applications of these gap forming protocols will utilize MutH, it is pertinent to note that the requirement of methyl-directed strand incision by MutH may be obviated by provision of a single-strand nick by some other means within the vicinity of the mispair, as described in Example 1, FIG. 5. A suitable means for

18

inducing such nicks in DNA is limited contact with a nuclease, Dnase I, for example; under conditions that are well known in the art, this approach creates nicks randomly throughout double-stranded DNA molecules at suitable intervals for allowing the mispair correction system to create single-stranded gaps in the vicinity of a mispair anywhere in the DNA.

It should be noted that in this embodiment of this method for localizing mismatch base pairs, "in the vicinity of" a base mispair is defined practically by the size of the single-strand gaps typically observed under above conditions, namely up to about one kbp from the mismatched base pair.

By "determining the location of the single-stranded gaps within the DNA duplex" entails the steps of: (i) Cleaving the DNA with at least one restriction endonuclease (either prior or subsequent to contact of the preparation with mismatch repair activities) followed by comparison of electrophoretic mobilities under denaturing conditions of the resulting modified DNA fragments with DNA restriction fragments not contacted with the defined mismatch correction system; or (ii) Cleaving the DNA with at least one restriction endonuclease and with a single-strand specific endonuclease, followed by comparison of the electrophoretic mobilities under native conditions of the resulting modified DNA fragments with DNA restriction fragments not contacted with the defined mismatch correction system. Suitable single-strand specific endonucleases include the S1 single-stranded specific nuclease, for example, or other functionally similar nucleases well known in the art. In the cases of either (i) or (ii), additional restriction mapping may be performed as needed to further localize any fragment modifications observed in initial application of the method, until, if desired, a restriction fragment of convenient size for direct sequence determination is obtained for direct comparisons of sequences of the two DNA molecules in the vicinity of the base sequence difference.

By "proteins of a mismatch repair system" are meant a protein that contains a GATC endonuclease, a mispair recognition protein, and proteins that participate in the activation of the GATC endonuclease.

By "divalent cation:" is meant a cofactor for the GATC endonucleases, e.g., $MgCl_2$.

By "endonucleolytic incision:" is meant cleavage of a DNA fragment containing a mismatched base pair at unmethylated or hemimethylated GATC sequences in the vicinity of a mismatch.

"Size fractionation by electrophoretic mobility under denaturing conditions" is a procedure well known by those skilled in the art. Gel Electrophoresis can either be conventional or pulse-field.

Modification of Mismatch Recognition Proteins and Uses

The present invention also includes forms of mispair recognition proteins which have been altered to provide means for modifying at least one strand of the DNA duplex in the vicinity of the bound mispair recognition protein.

In preferred embodiments of this aspect of the invention, the altered mispair recognition protein is the modified product of the mutS gene of *E. coli* or is another functionally homologous modified protein to which is attached an hydroxyl radical cleaving function; the altered mispair recognition protein may comprise only a segment of the native molecule containing the mispair recognition domain; the hydroxyl radical cleaving function is selected from the group

5,556,750

19

consisting of the altered mispair recognition protein wherein the hydroxyl radical cleaving function is selected from the group consisting of the 1,10-phenanthroline-copper complex, the EDTA iron complex, and the copper binding domain of serum albumin; the altered mispair recognition protein is the product of the *mutS* gene of *E. coli* or of another functionally homologous protein to which is attached attachment a DNA endonuclease activity capable of cleaving double-stranded DNA; the endonuclease activity is provided by the DNA cleavage domain of FokI endonuclease.

By "altered mispair recognition protein" is meant a mispair recognition protein that not only recognizes and binds to a base pair mismatch, but possess the ability to modify a strand of a DNA molecule containing such a mismatch.

Several methods for attaching an hydroxyl radical cleaving function to a DNA binding protein are known in the art. For example, lysyl residues may be modified by chemically attaching the 1,10-phenanthroline-copper complex to lysine residues, resulting in conversion of a DNA binding protein into a highly efficient site-specific nuclease that cleaved both DNA strands (in the presence of hydrogen peroxide as a coreactant) within the 20 base pair binding site of the protein, as determined by DNase I footprinting (C.-H. Chen and D. S. Sigman, 1987, *Science*, 237, 1197). Chemical attachment of an EDTA-iron complex to the amino terminus of another DNA binding protein similarly produced a sequence specific DNA cleaving protein that cut both strands of the target DNA within a few bases of recognition site of similar size (J. P. Sluka, et al., 1987, *Science*, 235, 777).

An alternate means for attaching the hydroxyl radical cleaving function to this same protein involved extension of the amino terminus with the three amino acids, Gly-Gly-His, which is consensus sequence for the copper-binding domain of serum albumin (D. P. Hack et al., 1988, *J. Am. Chem. Soc.*, 110, 7572-7574). This approach allows for preparation of such an artificial DNA cleaving protein directly by recombinant methods, or by direct synthesis using standard solid phase methods, when the peptide is sufficiently short as it was in this case (55 residues including the 3 added amino acids), thereby avoiding the need for an additional chemical modification step of the reagent which is both time consuming and difficult in large scale production. In contrast to the EDTA-iron complex, the particular peptide sequence constructed in this instance cleaved only one example out of four recognition sites in different sequence environments.

Nevertheless, one skilled in the art of protein engineering would appreciate that this general approach for converting a DNA binding protein into a DNA cleaving protein by attachment of an hydrogen radical cleavage function is widely applicable. Hence, DNA base mispair recognition proteins which normally only bind to DNA are modified to cleave DNA by attachment of an hydroxyl radical cleavage function, according to the practice of this aspect of this invention, without undue experimentation, by adjustment of appropriate variables taught in the art, particularly the chemical nature and length of the "spacer" between the protein and the metal binding site.

Additional altered forms of mispair recognition proteins that modify at least one strand of the DNA in a DNA:protein complex in the vicinity of the bound protein according to the present invention include proteins comprising the portions or "domains" of the unmodified base mispair recognition enzymes that are essential for binding to a DNA mispair. These essential DNA binding domains further comprise peptide sequences that are most highly conserved during

20

evolution; such conserved domains are evident, for example, in comparisons of the sequences of the *E. coli* MutS protein with functionally homologous proteins in *S. typhimurium* and other structurally similar proteins. Accordingly, peptide sequences of a DNA base mispair recognition protein that are protected from proteases by formation of specific complexes with mispairs in DNA and, in addition or in the alternative, are evolutionarily conserved, form the basis for a particularly preferred embodiment of this aspect of the present invention, since such peptides constitute less than half the mass of the intact protein and, therefore, are advantageous for production and, if necessary, for chemical modification to attach a cleavage function for conversion of the DNA binding protein into a DNA cleavage protein specific for sites of DNA base mispairs.

The DNA cleavage domain of FokI endonuclease has been defined (Li et al, 1992, *Proc. Natl. Acad. Sci. U.S.A.*, 89:4275).

Another embodiment of this aspect of the invention consists of a method for detecting and localizing a base pair mismatch within a DNA duplex, including the steps of contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs; contacting resultant duplex DNA molecules with an altered mispair recognition protein, under conditions such that the protein forms specific complexes With a mispair and thereby directs modification of at least one strand of the DNA in the resulting DNA protein complexes in the vicinity of the DNA protein complex, and determining the location of the modification of the DNA by a suitable analytic method.

In the detection and localization of a base pair mismatch method according to this embodiment which employs an altered mispair recognition protein, and the modification comprises double-stranded cleavage of the DNA within the vicinity of any base mispair wherein the "vicinity" substantially corresponds to the sequence of DNA protected by the binding of the protein to a base mispair, generally within about 20 base pairs. A single-strand specific nuclease, S1, for instance, may be used to augment cleavage by the modified base mispair recognition protein in the event that a single-strand bias is suspected in the cleavage of any DNAs with which the protein forms a specific complex. Alternatively, DNA's subject to cleavage by the modified mispair recognition protein may be analyzed by electrophoresis under denaturing conditions. Location of the modification is by suitable analytical methods known to those skilled in the art.

Methods Utilizing Mismatch Repair Systems to Detect A-G Base Pair Mismatches

In a preferred embodiment, a method for detecting and localizing A-G mispairs in a DNA duplex, includes the steps of contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs; contacting resultant duplex DNA molecules with a mispair recognition protein that recognizes A-G mispairs and an apurinic endonuclease or lyase under conditions such that in the presence of a mismatch an endonucleolytic incision is introduced in the duplex molecule, and determining the location of the incision by a suitable analytic method.

In preferred embodiments the A-G mispair recognition protein is the product of the *mutY* gene of *E. coli*; and the analytical method includes gel electrophoresis.

The present invention also comprises DNA mispair recognition protein that recognizes primarily A-G mispairs

5,556,750

21

without any apparent requirement for hemimethylation. One example of this protein is the product of the *mutY* gene of *E. coli*, is a glycosylase which specifically removes the adenine from an A-G mispair in a DNA duplex. The MutY protein has been purified to near homogeneity by virtue of its ability to restore A-G to C●G mismatch correction to cell-free extracts (K. G. Au et al., *Proc. Nat. Acad. Sci. U.S.A.*, 85, 9163, 1988) of a *mutS mutY* double mutant strain of *E. coli*, as described in Example 2, below. Its electrophoretic migration in the presence of dodecyl sulfate in consistent with a molecular weight of 36 kDa, and it apparently exists as a monomer in solution. MutY, an apurinic (AP) endonuclease, DNA polymerase I, and DNA ligase are sufficient to reconstitute MutY-dependent, A-G to C●G repair in vitro. A DNA strand that has been depurinated thusly by the MutY protein is susceptible to cleavage by any of several types of AP endonuclease or lyase (e.g., human AP endonuclease II) or by piperidine, under conditions that are well known in the art. The cleavage products are then analyzed by gel electrophoresis under denaturing conditions. Accordingly, this MutY protein is useful in a method for the specific detection and localization of A-G mispairs, according to the practice of the present invention, and hence identification of A●T to C●G or G●C to T●A mutations.

Sources of DNA Fragments to Be Analyzed

In another embodiment of the invention, DNA molecules are obtained from the following sources: different individuals of the same species, individuals of different species, individuals of different kingdoms, different tissue types, the same tissue type in different states of growth, different cell types, cells of the same type in different states of growth, and cells of the same origin in different stages of development, and cells of the same type that may have undergone differential somatic mutagenesis, e.g., one class of which may harbor per-cancerous mutation(s).

In a preferred embodiment, the DNA molecules comprise a probe sequence that has been at least partially characterized.

By "probe sequence that has been at least partially characterized" is meant a DNA molecule from any source that has been characterized by restriction mapping or sequence analysis, such techniques are known to those skilled in the art.

Kits Comprising a Mismatch Recognition Protein

Another aspect of the invention features assay kits designed to provide components to practice the methods of the invention.

In one aspect the invention features an assay kit for detecting a base pair mismatch in a DNA duplex. The kit comprises one or more of the following components: an aliquot of a mismatch recognition protein, an aliquot of control oligonucleotides, and an exonuclease.

In a preferred embodiment the mismatch recognition protein is the product of the *mutS* gene of *E. coli*.

By "control oligonucleotides" is meant oligonucleotides for assaying the binding of the mismatch repair protein to a base pair mismatch. One set of oligonucleotides are perfectly homologous (negative control) and thus are not bound by the mismatch recognition protein. Another set of oligonucleotides containing a base pair mismatch (positive control) and thus are bound by the mismatch recognition protein.

22

By "exonuclease" is meant enzymes possessing double-strand specific exonuclease activity, e.g., *E. coli* exonuclease III, RecBCD exonuclease, lambda exonuclease, and T7 gene 6 exonuclease.

Another aspect of the invention features an assay kit for detecting and localizing a base pair mismatch in a DNA duplex. The kit comprises one or more of the following components: an aliquot of all or part of a mismatch repair system, an aliquot of dideoxynucleoside triphosphates; and a single-strand specific endonuclease.

By "all or part of a mismatch repair system" is meant either the complete system which is capable of repairing a base pair mismatch, for example, the three *E. coli* proteins MutH, MutL, and MutS, DNA helicase II, single-strand binding protein, DNA polymerase III, exonuclease I, exonuclease VII or RecJ exonuclease, DNA ligase and ATP, or only the three proteins MutH, MutL, and MutS, along with ATP such that an endonucleolytic incision is made at a GATC site, with no subsequent repair reaction taking place.

In preferred embodiments the mismatch repair system includes: the products of the *E. coli* *mutH*, *mutL*, and *mutS* genes, or species variations thereof, DNA helicase II, single-strand DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, exonuclease VII or RecJ exonuclease, DNA ligase, and ATP, the mismatch repair system includes only the products of the *E. coli* *mutH*, *mutL*, and *mutS* genes, or species variations thereof, and ATP.

Another embodiment of the invention feature an assay kit for detecting and localizing a base pair mismatch in a DNA duplex comprising an aliquot of a modified mismatch recognition protein.

In a preferred embodiment the mismatch recognition protein is the product of the *mutS* gene of *E. coli*.

A further embodiment of this aspect of the invention features an assay kit for detecting and localizing an A-G mispair within a DNA duplex. The kit comprises one or more of the following components: an aliquot of an A-G mismatch recognition protein; and an aliquot of an apurinic endonuclease or lyase.

In a preferred embodiment the A-G mismatch recognition protein is the product of the *MutY* gene of *E. coli*.

Methods Utilizing Mismatch Repair Systems and Recombinase Proteins

In a further aspect, the invention features a method for eliminating DNA molecules containing one or more mismatches from a population of heterohybrid duplex DNA molecules formed by base pairing of single-stranded DNA molecules obtained from a first source and a second source. The method includes digesting genomic DNA from the first and the second source with a restriction endonuclease, methylating the DNA of one of the sources, denaturing the DNA from one or both sources, mixing the DNA molecules from the first and the second source in the presence of a recombinase protein, proteins of a mismatch repair system that modulate the recombinase protein, single-strand binding protein, and ATP under conditions such that DNA duplexes form in homologous regions of the DNA molecules from the first and the second source and the presence of a base pair mismatch results in regions that remain single-stranded, and removing molecules that contain single-stranded regions from the population.

By "heterohybrid" is meant a duplex DNA molecule that consists of base-paired strands originating from two different sources, such that one strand of the duplex is from one

5,556,750

23

source (first source) and the other strand is from another source (second source).

The "source" of DNA molecules designates the origin of the genomic DNA used in the method. The first and second sources are different, i.e., not from the same cell of the same individual.

By "restriction endonuclease" is meant an enzyme which recognizes specific sequences in double-stranded DNA and introduces breaks the phosphodiester backbone of both strands. For use in the current invention restriction endonucleases that digest genomic DNA or cDNA into fragments of approximately 4 to 20 kilobases are preferred.

By "methylating" is meant the process by which a methyl groups is attached to the adenine residue of the sequence "GATC". This reaction is carried by enzymes well known in the art, such as the DAM system of *E. coli*.

By "denaturing" is meant the process by which strands of duplex DNA molecules are no longer based paired by hydrogen bonding and are separated into single-stranded molecules. Methods of denaturation are well known to those skilled in the art and include thermal denaturation and alkaline denaturation.

By "recombinase protein" is meant a protein that catalyzes the formation of DNA duplex molecules. Such a molecule is capable of catalyzing the formation of duplex DNA molecules from complementary single-stranded molecules by renaturation or by catalyzing a strand transfer reaction between a single-stranded molecule and a double-stranded molecule. Examples of such a protein are the RecA proteins of *E. coli* and *S. typhimurium*.

By "proteins of a mismatch repair system that modulate the recombinase protein" are meant components of a system which recognizes and corrects base pairing errors in duplex DNA molecules and also influence the activity of a recombinase protein. For example, a mismatch recognition protein, e.g., MutS, and a protein that interacts with the mismatch repair protein, e.g., MutL, together inhibit duplex formation catalyzed by the recombinase protein in the presence of a base pair mismatch. Such modulation of the recombinase protein results in single-stranded regions downstream of the base pair mismatch.

In preferred embodiments, the recombinase protein is the *E. coli* RecA protein, the mismatch repair system is from *E. coli* and the components are the MutS and MutL proteins, the sources of DNA are different individuals of the same species, individuals of different species, individuals of different kingdoms, different tissue types, the same tissue type in different states of growth, different cell types, cells of the same type in different states of growth, cells of the same origin in different stages of development, and cells of the same origin that may have undergone differential somatic mutagenesis, the method of removing molecules containing single-stranded regions is by chromatography on benzoylated naphthoylated DEAE, the method of removing molecules containing single-stranded regions is by treatment with a single-strand specific nuclease.

The MutS, MutL protein, along with single-strand binding protein and ATP are involved in modulation of the *E. coli* RecA protein in catalyzing heteroduplex formation.

The method for removing molecules containing single-strands from double-stranded molecules by the use of chromatography with benzoylated naphthoylated DEAE is well known to those skilled in the art.

By "single strand specific nuclease" is meant an enzyme that specifically degrades single-stranded regions of DNA

24

molecules and do not degrade double stranded regions. Examples of such nucleases are: S1, mung bean, T7 gene 3 endonuclease and P1 nuclease.

In another aspect, the invention features a method for eliminating DNA molecules containing one or more mismatches from a population of heterohybrid duplex DNA molecules formed by a strand transfer reaction between duplex DNA molecules obtained from a first source and denatured DNA molecules from a second source. The method includes digesting genomic DNA from the first and the second source with a restriction endonuclease, methylating the DNA of one of the sources, denaturing the DNA from the second source, mixing the DNA molecules from the first and the second source in the presence of a protein which catalyzes strand transfer reactions, proteins of a mismatch repair system that modulate the protein with strand transfer activity, single strand binding protein, and ATP under conditions such that DNA heteroduplexes form in homologous regions of the DNA molecules from the first and the second source by strand transfer reaction and the presence of a base pair mismatch results in regions that remain single-stranded, and removing molecules that contain the single-stranded regions from the population.

By "strand transfer reaction is meant" a three strand reaction between duplex DNA from one source and single-stranded DNA from another source in which one strand of the duplex is displaced the by a single-stranded molecule.

By "a protein which catalyzes strand transfer reaction" is meant proteins such as: RecA, homologs of RecA, and proteins with branch migration enhancing activities such as RuvA, RuvB, RecG.

In preferred embodiments, the strand transferase protein is the *E. coli* RecA protein, the mismatch repair system is from *E. coli* and the components are the MutS and MutL proteins, the sources are different individuals of the same species, individuals of different species, individuals of different kingdoms, different tissue types, the same tissue type in different states of growth, different cell types, cells of the same type in different states of growth, and cells of the same origin in different stages of development, cells of the same origin that may have undergone differential somatic mutagenesis (e.g., normal as opposed to pre-tumor cells), a probe sequence that has been at least partially characterized, the method of removing molecules containing single-stranded regions is by chromatography on benzoylated naphthoylated DEAE, the method of removing molecules containing single-stranded regions is by treatment with a single strand specific nuclease.

Methods of Improving the Genomic Mismatch Scanning Technique

In another aspect the invention features the utilization of a recombinase or strand transferase and proteins of a mismatch repair system that modulate the recombinase or strand transferase, in the hybridization step of the genomic mismatch scanning technique. Formation of duplex molecules catalyzed by a recombinase or strand transferase protein which is modulated by components of a mismatch repair system, provide an additional selection step in the GMS method.

By "genomic mismatch scanning" is meant a technique to identify regions of genetic identity between two related individuals. Such a technique has been described by Nelson et al, 4 *Nature Genetics* 11, 1993.

In a further embodiment the invention features a method of genomic mismatch scanning such that heterohybrid DNA

5,556,750

25

molecules containing a base pair mismatch are removed, without the use of exonuclease III. The method comprises the steps of contacting a population of heterohybrid DNA molecules potentially containing base pair mismatches with all the components of a DNA mismatch repair system in the absence of dNTP's or in the presence of one or more dideoxy nucleoside triphosphates under conditions such that single-stranded gaps are generated in DNA fragments that contained a base pair mismatch and removing the molecules containing single-stranded gaps.

In preferred embodiments the DNA mismatch repair system is the *E. coli* methyl-directed mismatch repair system; removal of molecules containing single-stranded regions is by chromatography on benzoylated naphthoylated DEAE; removal of molecules containing single-stranded regions is by treatment with a single-strand specific nuclease.

In a further embodiment, the invention features another variation of the method of genomic mismatch scanning such that heterohybrid DNA molecules containing base pair mismatches are removed, without the use of exonuclease III. The method comprises the steps of contacting a population of heterohybrid DNA molecules potentially containing base pair mismatches with all the components of a DNA mismatch repair system and biotinylated nucleoside triphosphates under conditions such that biotinylated nucleotides are incorporated into DNA fragments that contained a base pair mismatch and, removing the molecules containing biotinylated molecules by binding to avidin.

Substitution with biotinylated nucleotides and binding of molecules that have incorporated these nucleotides are procedures well known to those skilled in the art. This procedure allows fractionation of a population of hybrid DNA molecules into two fractions: (i) A mismatch free fraction which fails to adhere to avidin; and (ii) A population that originally contained mispairs and which binds to avidin. The former can be utilized in the GMS procedure. The latter, avidin-bound class can be employed for other purposes. For example, when prepared using heterohybrid DNA produced by annealing DNA from two related haploid organisms the biotinylated sequences correspond to those DNA regions that vary genetically between the two organisms. Such sequences can thus be applied to determination of the molecular basis of genetic variation of organisms in question, e.g., pathogenic versus nonpathogenic microbial subspecies.

In a preferred embodiment the mismatch repair system is the methyl-directed mismatch repair system of *E. coli*.

In a further embodiment, the invention features a method of genomic mismatch scanning such that duplex DNA molecules are subject to exonuclease III digestion only after ligation into monomer circles.

By "ligation into monomer circles" is meant ligation of molecules under conditions of dilute concentration such that ends of the same molecule become ligated. Such a procedure is known to those skilled in the art. In these methods it is advantageous sometimes to separate molecules having mismatches from those which do not. By use of appropriate separation procedures both such populations of molecules can be selected.

Methods Applying Mismatch Repair Stems to Populations of Amplified Molecules

In another aspect, the invention features a method for correcting base pair mismatches in a population of DNA

26

duplexes that have been produced by enzymatic amplification potentially containing one or more base pair mismatches. The method includes contacting the population of DNA duplexes with a DNA methylase and a mismatch repair system such that base pair mismatches are corrected.

By "enzymatic amplification" is meant a reaction by which DNA molecules are amplified. Examples of such reactions include the polymerase chain reaction and reactions utilizing reverse transcription and subsequent DNA amplification of one or more expressed RNA sequences.

By "mismatch repair system" is meant a complete system such that base pair mismatches are detected and corrected.

In a preferred embodiment, the mismatch repair system is the methyl-directed mismatch repair system of *E. coli*. Components of the defined system capable of correcting mismatches include MutH, MutL, and MutS proteins, DNA helicase II, single-strand binding protein, DNA polymerase III holoenzyme, exonuclease I, exonuclease VII or RecJ, DNA ligase, ATP and four deoxynucleoside triphosphates.

In a further aspect, the invention features a method for removing DNA molecules containing one or more base pair mismatches in a population of molecules that have been produced by enzymatic amplification potentially containing one or more base pair mismatches. The method includes contacting a population of enzymatically amplified molecules with components of a mismatch repair system under conditions such that one or more components of the repair system form a specific complex with a base pair mismatch contained in a DNA duplex and removing DNA duplexes containing the complex from the population of duplex molecules.

By "complex" is meant the result of specific binding of at least one component of mismatch repair system to a base pair mismatch.

In a preferred embodiment, the mismatch repair system is the *E. coli* methyl-directed mismatch repair system, the component of the system is the MutS protein, the MutS protein is affixed to a solid support and removal of the DNA duplex containing the complex is by binding to this support.

Methods of attachment of proteins to solid support systems and use of those systems to perform chromatography so as to remove specific molecules are well known to those skilled in the art.

In another embodiment, the invention features a method for removing DNA molecules containing one or more base pair mismatches in a population of DNA duplexes that have been produced by enzymatic amplification, potentially containing one or more base pair mismatches. The method comprises the steps of contacting the population of DNA duplexes with components of a mismatch repair system under conditions such that an endonucleolytic incision is made on a newly synthesized strand of a DNA duplex molecule containing a base pair mismatch so that such a molecule cannot produce a full-sized product in a subsequent round of enzymatic amplification.

By "endonucleolytic cleavage" is meant cleavage on the unmethylated strand at a hemimethylate of GATC sequence by components of a mismatch repair system.

By "full sized product" is meant a molecule that includes the entire region of interest that is subject to amplification. Molecules that contain endonucleolytic cleavage cannot be amplified in subsequent rounds to produce full sized product and thus will be eliminated from the final amplified product population.

In a preferred embodiment the mismatch repair system is the methyl-directed mismatch repair system of *E. coli* and

5,556,750

27

the components are Muts, MutL, and MutH proteins, and ATP.

Methods to Remove from a Population Molecules Containing a Base Pair Mismatch

In a further embodiment the invention features a method for removing DNA duplex molecules containing base pair mismatches in a population of heteroduplex DNA molecules produced from different sources. The method comprises contacting the population of DNA duplex molecules potentially containing base pair mismatches with some or all components of a mismatch repair system under conditions such that the component or components form a complex with the DNA having a base pair mismatch, and not with a DNA duplex lacking a base pair mismatch, and removing DNA molecules containing the complex or the product of the complex.

By "product of the complex" is meant a DNA duplex that has incorporated biotinylated nucleotides.

By "some or all components of a mismatch repair system" is meant either a complete mismatch repair system such that the complete reaction is carried out or only the proteins of the system which specifically bind to the mismatch.

In preferred embodiments the mismatch repair system is the methyl-directed mismatch repair system of *E. coli*; some or all protein of the mismatch repair system have been affixed to a solid support and removal by adsorption; the complex interacts with other cellular proteins, and removal of the complex occurs through the interaction; and the conditions include the use of biotinylated nucleotides such that the nucleotides are incorporated into duplex molecules that contained a base pair mismatch and such duplexes are removed by binding to avidin.

By "some or all proteins" is meant, for example, *E. coli* proteins MutS, MutL, and MutH.

By "attached to a solid support" is meant a means, such as by fusion with glutathione transferase, by which a protein is attached to a solid support system and still remains functional.

By "adsorption" is meant specific binding to some or all of the proteins of the mismatch repair system affixed to a solid support so that separation from other molecules that do not bind to the solid support affixed proteins occurs.

By "interacts with other cellular proteins" is meant interaction between mismatch repair system protein or between those proteins and other proteins. For example, the interaction of MutS bound to a duplex DNA containing a mismatch with MutL or RecA.

Kits Containing a Mismatch Repair System

In a preferred embodiment, a kit for correcting base pair matches in duplex DNA molecules including one or more of the following components comprising the following purified components: an aliquot of *E. coli* MutH, MutL, and MutS proteins or species variations thereof, an aliquot of DNA helicase II, an aliquot of single-strand DNA binding protein, an aliquot of DNA polymerase III holoenzyme, an aliquot of exonuclease I, an aliquot of Exo VII or RecJ, an aliquot of DNA ligase, an aliquot of ATP, and an aliquot of four deoxynucleoside triphosphates.

A further embodiment of this aspect of this invention includes an assay kit for eliminating DNA molecules containing one or more base pairing mismatches from a population of heterohybrid duplex molecules formed by base

28

pairing of single-stranded DNA molecules obtained from a first and a second source comprising one or more of the following components, an aliquot of proteins of a mismatch repair system, and an aliquot of a recombinase protein.

By "proteins of a mismatch repair system" are meant proteins that modulate the activity of a recombinase protein.

In a preferred embodiment, the proteins of the mismatch correction system are the MutS and MutL proteins of *E. coli*.

Another aspect of the invention features an assay kit for removing DNA molecules containing one or more base pair mismatches comprising an aliquot of one or more proteins of a mismatch repair system that have been affixed to a column support.

In a preferred embodiment, the protein of the mismatch repair system is the MutS protein of *E. coli*.

Another aspect of the invention features a kit for fractionating a heteroduplex DNA population into two pools, one of which was mismatch-free at the beginning of the procedure, the second of which represents duplexes that contained mispaired bases at the beginning of the procedure. This kit is comprised of one or more of the following components: an aliquot of all components of complete mismatch repair system; an aliquot of biotinylated nucleotides; and an aliquot of avidin or an avidin-based support.

In a preferred embodiment, the mismatch repair system is from *E. coli* and consists of products of the mutH, mutL, and mutS genes, DNA helicase II, single-strand DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, exonuclease VII or RecJ exonuclease, DNA ligase, and ATP.

The following Examples are provided for further illustrating various aspects and embodiments of the present invention and are in no way intended to be limiting of the scope.

EXAMPLE 1

DNA Mismatch Correction in a Defined System

In order to address the biochemistry of methyl-directed mismatch correction, the reaction has been assayed in vitro using the type of substrate illustrated in FIG. 1. Application of this method to cell-free extracts of *E. coli* (A. L. Lu, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983) confirmed in vivo findings that methyl-directed repair requires the products of four mutator genes, mutH, mutL, mutS and uvrD (also called mutU), and also demonstrated a requirement for the *E. coli* single-strand DNA binding protein (SSB). The dependence of in vitro correction on mutH, mutL and mutS gene products has permitted isolation of these proteins in near homogeneous, biologically active forms. The MutS protein binds to mismatched DNA base pairs; the MutL protein binds to the MutS-heteroduplex complex (M. Grilley, K. M. Welsh, S.-S. Su, P. Modrich, *J. Biol. Chem.* 264, 1000, 1989); and the 25-kD MutH protein possesses a latent endonuclease that incises the unmethylated strand of a hemimethylated d(GATC) site (K. M. Welsh, A.-L. Lu, S. Clark, P. Modrich, *J. Biol. Chem.* 262, 15624, 1987), with activation of this activity depending on interaction of MutS and MutL with a heteroduplex in the presence of ATP (P. Modrich, *J. Biol. Chem.* 264, 6597, 1989). However, these three Mut proteins together with SSB and the DNA helicase II product of the uvrD (mutU) gene (I. D. Hickson, H. M. Arthur, D. Bramhill, P. T. Emmerson, *Mol. Gen. Genet.* 190, 265, 1983) are not sufficient to mediate methyl-directed repair. Below is described identification of the remaining required components and recon-

5,556,750

29

stitution of the reaction in a defined system.

Protein and cofactor requirements for mismatch correction. Methyl-directed mismatch correction occurs by an excision repair reaction in which as much as several kilobases of the unmethylated DNA strand is excised and resynthesized (A.-L. Lin, K. Welsh, S. Clark, S.-S. Su, P. Modrich, *Cold Spring Harbor Symp. Quant. Biol.* 49, 589, 1984). DNA polymerase I, an enzyme that functions in a number of DNA repair pathways, does not contribute in a major way to methyl-directed correction since extracts from a polA deletion strain exhibit normal levels of activity. However extracts derived from a dnaZ^{ts} strain are temperature sensitive for methyl-directed repair in vitro (Table 1).

TABLE 1

Requirement for τ and γ Subunits of DNA Polymerase III Holoenzyme in Mismatch Repair			
Extract genotype	DNA Pol III addition (ng)	Mismatch Correction Activity (fmol/h/mg) Extract preincubation 42°	ratio (42°/34°)
dnaZ ^{ts}	—	8	910.09
	57 ng	75	1600.47
dnaZ ⁺	—	150	1600.94
	57 ng	160	1601.0

Extracts from strains AX727 (lac thi str^R dnaZ20-16) and AX729 (as AX727 except pure dnaZ⁺) were prepared as described (A.-L. Lin, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983). Samples (110 μ g of protein) were mixed with 0.8 μ l of 1M KCl and water to yield a volume of 7.2 μ l, and preincubated at 42° or 34° C. for 2.5 minutes. All heated samples were then placed at 34° C. and supplemented with 2.2 μ l of a solution containing 0.1 μ g (24 fmol) of hemimethylated G-T heteroduplex DNA, 16 ng of MutL protein, 50 ng of MutS protein, and buffer and nucleotide components of the mismatch correction assay (A.-L. Lu, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983). DNA polymerase III holoenzyme (57 ng in 0.6 μ l) or enzyme buffer was then added, and incubation at 34° C. was continued for 60 min. Heated extracts were supplemented with purified MutL and MutS proteins because these components are labile at 42° C. Activity measurements reflect the correction of heteroduplex sites.

The dnaZ gene encodes the τ and γ subunits of DNA polymerase III holoenzyme (M. Kodaira, S. B. Biswas, A. Kornberg, *Mol. Gen. Genet.* 192, 80, 1983; D. A. Mullin, C. L. Woldringh, J. M. Henson, J. R. Walker, *Mol. Gen. Genet.* 192, 73, 1983), and mismatch correction activity is largely restored to heated extracts of the temperature-sensitive mutant strain by addition of purified polymerase III holoenzyme. Since DNA polymerase III holoenzyme is highly processive, incorporating thousands of nucleotides per DNA binding event, the involvement of this activity is consistent with the large repair tracts associated with the methyl-directed reaction.

Additional data indicate that purified MutH, MutL, and MutS proteins, DNA helicase II, SSB, and DNA polymerase III holoenzyme support methyl-directed mismatch correction, but this reaction is inhibited by DNA ligase, an enzyme that is shown below to be required to restore covalent continuity to the repaired strand. This observation led to isolation of a 55-kD stimulatory protein that obviates ligase inhibition. The molecular weight and N-terminal sequence of this protein indicated identity to exonuclease I (G. J. Phillips and S. R. Kushner, *J. Biol. Chem.* 262, 455, 1987), and homogeneous exonuclease I readily substitutes for the

30

55-kD stimulatory activity (Table 2). Thus, exonuclease I and the six activities mentioned above mediate efficient methyl-directed mismatch correction in the presence of ligase to yield product molecules in which both DNA strands are covalently continuous.

TABLE 2

Stimulation of in vitro Methyl-Directed Correction by Exonuclease I.	
Protein added	Mismatch correction (fmol/20 min)
None	1
55-kD protein	18
Exonuclease I	18

Reactions (10 μ l) contained 0.05M HEPES (potassium salt, pH 8.0), 0.02M KCl, 6 mM MgCl₂, bovine serum albumin (0.05 mg/ml), 1 mM dithiothreitol, 2 mM ATP, 100 μ M (each) dATP, dCTP, dGTP, and dTTP, 25 μ M β -NAD⁺, 0.1 μ g of hemimethylated, covalently closed G-T heteroduplex DNA (FIG. 1, methylation on c strand, 24 fmol), 0.26 ng of MutH (K. M. Welsh, A.-L. Lin, S. Clark, P. Modrich, *J. Biol. Chem.* 262, 15624, 1987), 17 ng of MutL (M. Grilley, K. R. Welsh, S.-S. Su, P. Modrich, *J. Biol. Chem.* 264, 1000, 1989), 35 ng of MutS (S.-S. Sin and P. Modrich, *Proc. Natl. Acad. Sci. USA* 83, 5057, 1986), 200 ng of SSB (T. R. Lohman, J. R. Green, R. S. Beyer, *Biochemistry* 25, 21, 1986; U.S. Biochemical Corp.), 10 ng of DNA helicase II (K. Kumura and M. Sekiguchi, *J. Biol. Chem.* 259, 1560, 1984), 20 mg of *E. coli* DNA ligase (U.S. Biochemical Corp.), 95 ng of DNA polymerase III holoenzyme (C. McHenry and A. Kornberg, *J. Biol. Chem.* 252, 6478, 1977), and 1 ng of 55-kD protein or exonuclease I (U.S. Biochemical Corp.) as indicated. Reactions were incubated at 37° C. for 20 minutes, quenched at 55° C. for 10 minutes, chilled on ice, and then digested with Xho I or Hind III endonuclease to monitor correction. Repair of the G-T mismatch yielded a only the G-C containing, Xho I-sensitive product.

The requirements for repair of a covalently closed G-T heteroduplex (FIG. 1) are summarized in Table 3 (Closed circular). No detectable repair was observed in the absence of MutH, MutL, or MutS proteins or in the absence of DNA polymerase III holoenzyme, and omission of SSB or exonuclease I reduced activity by 85 to 90 percent.

TABLE 3

Protein and Cofactor Requirements for Mismatch Correction in a Defined System.		
Reaction conditions	Mismatch correction (fmol/20 min)	
	Closed Circular Heteroduplex	Open Circular Heteroduplex
Complete	15	17 (No MutH, No ligase)
minus MutH	<1	—
minus MutL	<1	<1
minus MutS	<1	<1
minus DNA polymerase III holoenzyme	<1	<1
minus SSB	2	1.4
minus exonuclease I	2	<1
minus DNA helicase II	16	15
minus helicase II,	<1	<1
plus immune serum	14	NT
minus helicase II,		
plus pre-immune serum		

5,556,750

31

TABLE 3-continued

Reaction conditions	Mismatch correction (fmol/20 min)	
	Closed Circular Heteroduplex	Open Circular Heteroduplex
minus Ligase/NAD ⁺	14	NT
minus MgCl ₂	<1	NT
minus ATP	<1	NT
minus dNTP's	<1	NT

Reactions utilizing covalently closed G-T heteroduplex (modification on c strand) were performed as described in the legend to Table 2 except that 1.8 ng of exonuclease I was used. Repair of open circular DNA was performed in a similar manner except that *RutH*, DNA ligase, and β -NAD⁺ were omitted from all reactions, and the hemimethylated G-T heteroduplex (modification on c strand) had been incised with *MutH* protein as described in the legend to FIG. 4. When present, rabbit antiserum to helicase II or preimmune serum (5 μ g protein) was incubated at 0° C. for 20 minutes with reaction mixtures lacking MgCl₂; the cofactor was then added and the assay was performed as above. Although not shown, antiserum inhibition was reversed by the subsequent addition of more helicase II. With the exception of the DNA polymerase III preparation, which contained about 15% by weight DNA helicase II (text) the purity of individual protein fractions was >95%. NT—not tested.

These findings are in accord with previous conclusions concerning requirements of the methyl-directed reaction. However, in contrast to observations *in vivo* and in crude extracts indicating a requirement for the *uvrD* product, the reconstituted reaction proceeded readily in the absence of the added DNA helicase II (Table 2). Nevertheless, the reaction was abolished by antiserum to homogeneous helicase II, suggesting a requirement for this activity and that it might be present as a contaminant in one of the other proteins. Analysis of these preparations for their ability to restore mismatch repair to an extract derived from a *uvrD* (*mutU*) mutant and for the physical presence of helicase II by immunoblot assay revealed that the DNA polymerase III holoenzyme preparation contained sufficient helicase II (13 to 15 percent of total protein by weight) to account for the levels of mismatch correction observed in the defined system. Similar results were obtained with holoenzyme preparations obtained from two other laboratories. The purified system therefore requires all the proteins that have been previously implicated in methyl-directed repair.

The rate of correction of the closed circular heteroduplex was unaffected by omission of DNA ligase (Table 3), but the presence of this activity results in production of a covalently closed product. Incubation of a hemimethylated, supercoiled G-T heteroduplex with all seven proteins required for correction in the presence of DNA ligase resulted in extensive formation of covalently closed, relaxed, circular molecules. Production of the relaxed DNA was dependent on *MutS* (FIG. 2) and *MutL* proteins, and the generation of this species was associated with heteroduplex repair (FIG. 2). Correction also occurred in the absence of ligase, but in this case repair products were open circular molecules, the formation of which depended on the presence of *MutS* (FIG. 2). Since *MutS* has no known endonuclease activity but does recognize mispairs, it is inferred that open circular mol-

32

ecules are the immediate product of a mismatch-provoked-excision repair process. Ligase closure of the strand break(s) present in this species would yield the covalently closed, relaxed circular product observed with the complete system.

The set of purified activities identified here as being important in methyl-directed repair support efficient correction. In the experiments summarized in Table 3, the individual proteins were used at the concentrations estimated to be present in the standard crude extract assay for correction as calculated from known specific activity determinations. Under such conditions the rate and extent of mismatch repair in the purified system are essentially identical to those observed in cell-free extracts.

DNA sites involved in repair by the purified system. The single d(GATC) sequence within the G-T heteroduplex shown in FIG. 1 is located 1024 base pairs from the mispair. Despite the distance separating these two sites, correction of the mismatch by the purified system responded to the state of modification of the d(GATC) sequence as well as its presence within the heteroduplex (FIG. 3). A substrate bearing d(GATC) methylation on both DNA strands did not support mismatch repair nor did a related heteroduplex in which the d(GATC) sequence was replaced by d(GATT). However, each of the two hemimethylated heteroduplexes were subject to strand-specific correction, with repair in each case being restricted to the unmodified DNA strand. With a heteroduplex in which neither strand was methylated, some molecules were corrected on one strand, and some were corrected on the other. As can be seen, the hemimethylated heteroduplex bearing methylation on the complementary DNA strand was a better substrate than the alternative configuration in which modification was on the viral strand, with a similar preference for repair of the viral strand being evident with the substrate that was unmethylated on either strand. This set of responses of the purified system to the presence and state of modification of d(GATC) sites reproduce effects previously documented *in vivo* and in crude extract experiments (R. S. Lahue, S.-S. Su, P. Modrich, *Proc. Natl. Acad. Sci. USA* 84, 1482, 1987).

TABLE 4

Heteroduplex	Mark- ers	Correction Efficiencies for Different Mismatches.			
		C ⁺ V ⁻		C ⁻ V ⁻	
		Rate	Bias	Rate	Bias
C 5'-CTCGA G AGCTT V 3'-GAGCT T TCGAA	Xho I Hind III	1.2	>18	0.38	>5
C 5'-OCTCGA G AGCTG V 3'-GAGCT G TCGAC	Xho I Pvu II	1.1	>17	0.38	>6
C 5'-ATCGA T AGCTT V 3'-TAGCT T TCGAA	Cla I Hind III	1.0	>16	0.24	3
C 5'-ATCGA A AGCTT V 3'-TAGCT A TCGAA	Hind III Cla I	0.88	>20	0.20	>7
C 5'-CTCGA A AGCTT V 3'-GAGCT C TCGAA	Hind III Xho I	0.61	17	0.28	>5
C 5'-GTCTGA C AGCTT V 3'-CAGCT T TCGAA	Sai I Hind III	0.60	12	0.23	>4
C 5'-GTCTGA A AGCTT V 3'-CAGCT T TCGAA	Hind III Sai I	0.44	>13	0.21	5
C 5'-CTCGA C AGCTG V 3'-GAGCT C TCGAC	Pvu II Xho I	0.04	NS	<0.04	NS

5,556,750

33

TABLE 4-continued

Heteroduplex	Mark- ers	Correction Efficiencies for Different Mismatches.			
		C ⁺ V ⁻		C ⁻ V ⁺	
		Rate	Bias	Rate	Bias

Table 4 (Continued) Correction of the eight possible base—base mispairs was tested with the set of covalently closed heteroduplexes described previously including the G—T substrate shown in FIG. 1. With the exception of the mispair and the variations shown at the fifth position on either side, all heteroduplexes were identical in sequence. Each DNA was tested in both hemimethylated configurations under complete reaction conditions (Table 3, closed circular heteroduplex) except the samples were removed at 5-minute intervals over a 20 minute period in order to obtain initial rates (fmol/min). c and v refer to complementary and viral DNA strands, and Bias indicates the relative efficiency of mismatch repair occurring on the two DNA strands (ratio of unmethylated to methylated) as determined 60 minutes after the reaction was started. NS — not significant. With the exception of the C—C heteroduplexes, repair in the absence of MutS protein was less than 20% (in most cases < 10%) of that observed in its presence (not shown).

The efficiency of repair by the methyl-directed pathway depends not only on the nature of the mispair, but also on the sequence environment in which the mismatch is embedded (P. Modrich, *Ann. Rev. Biochem.* 56, 435, 1987). To assess the mismatch specificity of the purified system under conditions where sequence effects are minimized, a set of heteroduplexes were used in which the location and immediate sequence environment of each mispair are essentially identical (S.-S. Su, R. S. Lahue, K. G. Au, P. Modrich, *J. Biol. Chem.* 263, 6829, 1988). This analysis (Table 4) showed that the purified system is able to recognize and repair in a methyl-directed manner seven of the eight possible base—base mismatches, with C—C being the only mispair that was not subject to significant correction. Table 3 also shows that the seven corrected mismatches were not repaired with equal efficiency and that in the case of each heteroduplex, the hemimethylated configuration modified on the complementary DNA strand was a better substrate than the other configuration in which the methyl group was on the viral strand. These findings are in good agreement with patterns of repair observed with this set of heteroduplexes in *E. coli* extracts (Although the patterns of substrate activity observed in extracts and in the purified system are qualitatively identical, the magnitude of variation observed differs for the two systems. Hemimethylated heteroduplexes modified on the complementary DNA strand are better substrates in both systems, but in extracts such molecules are repaired at about twice the rate of molecules methylated on the viral strand. In the purified system these relative rates differ by factors of 2 to 4. A similar effect may also exist with respect to mismatch preference within a given hemimethylated family. Although neither system repairs C—C, the rates of repair of other mismatches vary by a factors of 1.5 to 2 in extracts but by factors of 2 to 3 in the defined system.).

Strand-specific repair directed by a DNA strand break. Early experiments on methyl-directed repair in *E. coli* extracts led to the proposal that the strand-specificity of the reaction resulted from endonucleolytic incision of an unmethylated DNA strand at a d(GATC) sequence. This idea was supported by the finding that purified MutH protein has an associated, but extremely weak d(GATC) endonuclease that is activated in a mismatch-dependent manner in a reaction requiring MutL, MutS, and ATP. The purified system has been used to explore this effect more completely.

The two hemimethylated forms of the G—T heteroduplex shown in FIG. 1 were incised using high concentrations of purified MutH protein to cleave the unmethylated DNA strand at the d(GATC) sequence (>>pGpApTpC). After

34

removal of the protein, these open circular heteroduplexes were tested as substrates for the purified system in the absence of DNA ligase. Both open circular species were corrected in a strand-specific manner and at rates similar to those for the corresponding covalently closed heteroduplexes (FIG. 4). As observed with closed circular heteroduplexes, repair of the MutH-cleaved molecules required MutL, MutS, SSB, DNA polymerase III holoenzyme, and DNA helicase II (FIG. 4 and open circle entries of Table 2), but in contrast to the behavior of the closed circular substrates, repair of the mismatch within the open circular molecules occurred readily in the absence of MutH protein. Thus prior incision of the unmethylated strand of a d(GATC) site can bypass the requirement for MutH protein in strand-specific mismatch correction.

The nature of the MutH-independent repair was examined further to assess the effect of ligase on the reaction and to determine whether a strand break at a sequence other than d(GATC) can direct correction in the absence of MutH protein (FIG. 5). As mentioned above, a covalently closed G—T heteroduplex that lacks a d(GATC) sequence is not subject to repair by the purified system in the presence (FIG. 3) or absence of DNA ligase. However, the presence of one strand-specific, site-specific break is sufficient to render this heteroduplex a substrate for the purified system in the absence of ligase and Ruth protein (FIG. 5). Repair of this open circular heteroduplex was limited to the incised, complementary DNA strand, required presence of MutL and MutS proteins, DNA polymerase III, and SSB, and correction of the molecule was as efficient as that observed with the hemimethylated heteroduplex that had been cleaved by MutH at the d(GATC) sequence within the complementary strand. Although the presence of a strand break is sufficient to permit strand-specific correction of a heteroduplex in the absence of MutH and ligase, the presence of the latter activity inhibited repair not only on the heteroduplex lacking a d(GATC) sequence but also on both hemimethylated molecules that had been previously incised with MutH protein (FIG. 5). This inhibition by ligase was circumvented by the presence of MutH protein, but only if the Substrate contained a d(GATC) sequence, with this effect being demonstrable when both types of heteroduplex were present in the same reaction (FIG. 5, last column). This finding proves that MutH protein recognizes d(GATC) sites and is consistent with the view that the function of this protein in mismatch correction is the incision of the unmethylated strand at this sequence.

EXAMPLE 2

Purification of MutY Protein

Purification of MutY Protein *E. coli* RK1517 was grown at 37° C. in 170 liters of L broth containing 2.5 mM KH₂PO₄, 7.5 mM Na₂HPO₄ (culture, pH=7.4) and 1% glucose. The culture was grown to an A₅₉₀ of 4, chilled to 10° C. and cells were harvested by continuous flow centrifugation. Cell paste was stored at 70° C. A summary of the MutY purification is presented in Table 1. Fractionation procedures were performed at 0°–4° C., centrifugation was at 13,000×g, and glycerol concentrations are expressed as volume percent.

Frozen cell paste (290 g) was thawed at 4° C., resuspended in 900 ml of 0.05M Tris-HCl (pH 7.5), 0.1M NaCl, 1 mM dithiothreitol, 0.1 mM EDTA, and cells were disrupted by sonication. After clarification by centrifugation for 1 hr, the lysate (Fraction I, 970 ml) was treated with 185 ml

5,556,750

35

of 25% streptomycin sulfate (wt/vol in 0.05M Tris-HCl (pH 7.5), 0.1M NaCl, 1 mM dithiothreitol, 0.1 mM EDTA) which was added slowly with stirring. After 30 min of additional stirring, the solution was centrifuged for 1 h, and the supernatant (1120 ml) was treated with 252 g of solid ammonium sulfate which was added slowly with stirring. After 30 min. of additional stirring, the precipitate was collected by centrifugation for 1 h, resuspended to a final volume of 41 ml in 0.02M potassium phosphate (pH 7.5), 0.1 mM EDTA, 10% (vol/vol) glycerol, 1 mM dithiothreitol, and dialyzed against two 2 l portions of 0.02M potassium phosphate (pH 7.5), 0.1M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol (2 h per change). The dialyzed material was clarified by centrifugation for 10 min to yield Fraction II (45 ml).

Fraction II was diluted 10-fold into 0.02 M potassium phosphate (pH 7.5), 0.1M EDTA, 1 mM dithiothreitol, 10% glycerol so that the conductivity of the diluted solution was comparable to that of the dilution buffer containing 0.1M KCl. The solution was performed on small aliquots of Fraction II, and diluted samples were immediately loaded at 1 ml/min onto a 14.7 cm \times 12.6 cm² phosphocellulose column equilibrated with 0.02 M potassium phosphate (pH 7.5), 0.1M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. The column was washed with 400 ml of equilibration buffer, and developed with a 2 liter linear gradient of KCl (0.1 to 1.0M) in 0.02M potassium phosphate (pH 7.5), 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. Fractions containing MutY activity, which eluted at about 0.4M KCl, were pooled (Fraction III, 169 ml).

Fraction III was dialyzed against two 500 ml portions of 5 mM potassium phosphate (pH 7.5), 0.05M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol (2 h per change) until the conductivity was comparable to that of the dialysis buffer. After clarification by centrifugation for 10 min, the solution was loaded at 0.5 ml/min onto a 21 cm \times 2.84 cm² hydroxylapatite column equilibrated with 5 mM potassium phosphate, pH 7.5, 0.05 M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. After washing with 130 ml of equilibration buffer, the column was eluted with a 600 ml linear gradient of potassium phosphate (5 mM to 0.4M, pH 7.5) containing 0.05M KCl, 1 mM dithiothreitol, 10% glycerol. Fractions eluting from the column were supplemented with EDTA to 0.1 mM. Peak fractions containing 60% of the total recovered activity, which eluted at about 0.1M potassium phosphate, were pooled (Fraction IV, 24 ml). The remaining side fractions contained impurities which could not be resolved from MutY by MonoS chromatography.

Fraction IV was diluted by addition of an equal volume of 0.1 mM EDTA, 1 dithiothreitol, 10% glycerol. After clarification by centrifugation for 15 min, diluted Fraction IV was loaded at 0.75 ml/min onto a Pharmacia HR 5/5MonoS FPLC column that was equilibrated with 0.05M sodium phosphate (pH 7.5), 0.1M NaCl, 0.1 mM EDTA, 0.5 mM dithiothreitol, 10% glycerol.

36

TABLE I

Purification of MutY protein from 290 g of <i>E. coli</i> RK1517				
Fraction	Step	Total Protein mg	Specific Activity units/mg	Yield Percent
I	Extract	10,900	40	(100)
II	Ammonium sulfate	1,350	272	84
III	Phosphocellulose	66	10,800	160
IV	Hydroxylapatite	1.4	136,000	44
V	MonoS	0.16	480,000	18

Specific A.G to C—G mismatch correction in cell-free extracts was determined as described previously (Au et al. 1988), except that ATP and glutathione were omitted from the reaction and incubation was for 30 min instead of 1 h. For complementation assays, each 0.01 ml reaction contained RK1517-Y33 extract (mutS mutY) at a concentration of 10 mg/ml protein. One unit of MutY activity is defined as the amount required to convert 1 fmol of A.G mismatch to C—G base pair per h under complementation conditions.

The column was washed at 0.5 ml/min with 17 ml of equilibration buffer and developed at 05 ml/min with a 20 ml linear gradient of NaCl (0.1 to 0.4M) in 0.05M sodium phosphate (pH 7.5), 0.1 mM EDTA, 0.5 mM dithiothreitol, 10% glycerol. Fractions with MutY activity, which eluted at approximately 0.2M NaCl, were pooled (Fraction V, 2.6 ml). Fraction V was divided into small aliquots and stored at -70° C.

Assay for MutY-dependent, A●G-specific Glycosylase

DNA restriction fragments were labeled at either the 3' or 5' ends with ³²P. Glycosylase activity was then determined in 0.01 ml reactions containing 10 ng end-labeled DNA fragments, 0.02M Tris-HCl, pH 7.6, 1 mM EDTA, 0.05 mg/ml bovine serum albumin, and 2.7 ng MutY. After incubation at 37° C. for 30 min, the reaction mixture was treated with 2.5 \times 10⁻³ units of HeLa AP endonuclease II in the presence of 11 mM MgCl₂ and 0.005% Triton X-100 for 10 min at 37° C. Reactions were quenched by the addition of an equal volume of 80% formamide, 0.025% xylene cyanol, 0.025% bromphenol blue, heated to 80° C. for 2 min, and the products analyzed on an 8% sequencing gel. Control reactions contained either no MutY, no A●G mismatch or no AP endonuclease II.

Strand cleavage at the AP site generated by MutY could also be accomplished by treatment with piperidine instead of treatment with AP endonuclease II. After incubation for 30 min. at 37° C. with MutY as described above, the reaction mixture was precipitated with ethanol in the presence of carrier tRNA, then resuspended in 1M piperidine and heated at 90° C. for 30 min. After two additional ethanol precipitations, changing tubes each time, the pellet was resuspended in a minimum volume of water to which was added an equal volume of 80% formamide, 0.025% xylene cyanol, 0.025% bromphenol blue. The products were then analyzed on an 8% sequencing gel.

EXAMPLE 3

Genetic Mapping Point Mutations in the Human Genome

The full novelty and utility of the present invention may be further appreciated by reference to the following brief description of selected specific embodiments which advantageously employ various preferred forms of the invention

5,556,750

37

as applied to a common problem in genetic mapping of point mutations in the human genome. In the course of constructing gene linkage maps, for example, it is frequently desirable to compare the sequence of a cloned DNA fragment with homologous sequences in DNA extracted from a human tissue sample. Substantially all base pairs in the entire homologous sequence of the cloned DNA fragment are compared to those of the human tissue DNA, most advantageously in a single test according to the present invention, merely by contacting both strands of the human tissue DNA molecule with both radiolabeled complementary strands of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting DNA duplexes with the *E. coli* MutS protein that recognizes substantially all base pair mismatches under conditions such that the protein forms specific complexes with its cognate mispairs, and detecting the resulting DNA:protein complexes by contacting the complexes with a membranous nitrocellulose filter under conditions such that protein:DNA complexes are retained while DNA not complexed with protein is not retained, and measuring the amount of DNA in the retained complexes by a standard radiological methods or by utilizing any of the other methods of the invention; e.g., altered electrophoretic mobility, or detection by use of antibodies.

If the above detection test indicates the presence of sequence differences between the human tissue DNA and the cloned DNA and localization is required, or, in the alternative, if such differences are suspected and localization as well as detection of them is desired in a first analysis, the another method of this invention may be applied for these purposes. An embodiment of this aspect of the invention that may be most advantageously employed comprises the steps of contacting both strands of the human tissue DNA molecule with both radiolabeled complementary strands of the second DNA molecule (usually without separation from the cloning vector DNA) under conditions such that base pairing occurs, contacting the resulting DNA duplexes with MutHLS to produce a GATC cleavage reaction or a modified form of MutS protein of *E. coli* to which is attached an hydroxyl radical cleaving function under conditions such that the radical cleaving function cleaves both strands of the DNA within about 20 base pairs of substantially all DNA base mispairs. In the absence of any DNA base mispairs in the DNA duplexes comprising complementary strands of the human tissue and cloned DNAs, no DNA fragments smaller than the cloned DNA (plus vector DNA, if still attached) would be detected. Determination of the location of any double-stranded DNA cleavages by the modified MutS protein to within a few kbp or less of some restriction enzyme cleavage site within the cloned DNA is determined by standard restriction enzyme mapping approaches. If greater precision in localization and identification of a single base difference is desired, sequencing could be confined to those particular fragments of cloned DNA that span at least one base sequence difference localized by this method and are cleaved by a restriction enzyme at the most convenient distance of those sequence differences for direct sequencing.

The examples herein can be changed to make use of other methods of separation to identify mismatches, such as a filter-binding assay, as well as the nicking reaction with MutS and MutL. While large (at least 20 kbp) or small DNA molecules can be used in these methods those of between 1–10 kbp are preferred.

38

EXAMPLE 4

DNA Mismatch Detection Kit

Kit contains MutS protein, dilution buffer, annealing buffer, reagents to generate complementary and mismatched control duplexes and filter binding protocol. It can be used to detect single-base mismatches in oligonucleotides.

MutS kit components:

MutS protein in storage buffer: 50 mM HEPES pH7.2, 100 mM KCl, 1 mM EDTA, 1 mM DTT;

MutS1: 16 mer oligonucleotide GATCCGTCGACCT-GCA (all such oligonucleotides are written 5' to 3' herein) in water (2 μ M);

MutS2: 16 mer oligonucleotide TGCAGGTCGACG-GATC 1 μ M in annealing buffer 1 μ M: 20 mM Tris/HCl pH 7.6, 5 mM, MgCl₂, 0.1 mM DTT, 0.01 mM EDTA;

MutS3: 16 mer oligonucleotide TGCAGGTTGACG-GATC 1 μ M in annealing buffer;

Assay buffer/annealing buffer/wash buffer, 20 mM Tris/HCl pH 7.6, 5 mM MgCl₂, 0.1 mM DTT, 0.01 mM EDTA;

Protein storage/dilution buffer: 50 mM HEPES pH 7.2, 100 mM KCl, 1 mM EDTA, 1 mM DTT.

The DNA mismatch detection kit contains three 16-mer oligonucleotides labeled MUTS1, MUTS2, and MUTS3 for testing the performance of MutS protein. When MUTS1 and MUTS2 are annealed, a perfectly matched duplex results. When MUTS1 and MUTS3 are annealed, a duplex containing a single G–T mismatch results. These serve as control substrates for MutS binding.

Kinase Labeling of MUTS1 Oligonucleotide

This protocol uses half the amount of oligonucleotide contained in the kit. To a microcentrifuge tube on ice add the following:

MUTS1 Oligonucleotide (2 μ M)	15 μ l (30 pmoles)
10X T4 Polynucleotide Kinase Buffer	3 μ l
³² P-ATP (3000 Ci/mmol)	1 μ l
ATP (10 μ M)	2.5 μ l
Sterile dH ₂ O	7.5 μ l
T4 Polynucleotide Kinase (30 units/ μ l)	1 μ l (30 units)

Incubate the reaction mixture for 10 min at 37° C. Then incubate 10 min at 70° C. Spot two independent 1 μ l aliquots of the mixture on a SureCheck TLC plate and also spot a dilution of ³²P-ATP (1:30 in water) in a separate lane and run with the elution mixture. Expose the developed plate to X-ray film for 5 min. Scrape all radioactive spots from both experimental lanes of the plate and count them in a liquid scintillation counter to determine the % incorporation of label. This value is typically 40–60%. If a significant labeled ATP spot is present in the kinase reaction lanes on the plate, the labeled oligonucleotide must be purified before use (TLC or gel), since ³²P-ATP will contribute to background in the filter binding assay. In our experience, this is usually not necessary.

Keep in mind that the MUTS1 oligo stock is 2 pmol/ μ l and that the final concentration should be 1 pmol/ μ l. It is critical that this final concentration be as exact as possible, since the concentration determines the amount of MUTS1 in the next (annealing) step and hence, the amount of DNA available for binding by the protein.

5,556,750

39

Annealing Reactions

Two separate reactions are carried out: MUTS1/MUTS2 and MUTS1/MUTS3. In both cases, the ^{32}P -labeled MUTS1 from Step 1 is used.

Complementary		Mismatched	
MUTS1 (kinased)	14 μl = 14 pmol	MUTS1 (kinased)	14 μl = 14 pmols
MUTS2 (1 μM)	28 μl = 28 pmol	MUTS3 (1 μM)	28 μl = 28 pmols
annealing buffer	28 μl	annealing buffer	28 μl
	70 μl		70 μl

1. Heat each mixture for 10 min at 70° C.
2. Incubate for 30 min at room temperature.
3. Hold on ice until ready to use.

The molar ratio of MUTS2/MUTS1 and MUTS3/MUTS1 is 2:1 in the above reactions and this should be maintained for optimal results. Lowering the ratio of unlabeled to labeled strand may lead to very high background in the filter binding assay, presumably caused by sticking of labeled ssDNA to nitrocellulose.

Assay of MutS Binding by the Gel Shift Method

The binding of MutS to mismatches can be assessed using the technique of Gel Shift Mobility Assay (GSMA), a useful tool to identify protein-DNA interactions which may regulate gene expression. Below is a protocol for performing GSMA on the MUTS1/MUTS3 mismatched duplex contained in the mismatch detection kit. Optimum conditions may vary depending on the particular mismatch being detected or the length of the oligonucleotide.

All binding reactions should be carried out on ice. The total binding reaction volume is 10 μl . Add 4 μl of a MutS protein dilution (prepared using dilution buffer in the kit) containing 0.5–5 pmols (0.125–1.25 units) of MutS protein (1 pmol=97 ng) to 6 μl =1.2 pmols of ^{32}P -labeled MUTS1/MUTS3 heteroduplex. Also add comparable amounts of MutS protein to labeled MUTS1/MUTS2 matched duplex to serve as a control. A control incubation consisting only of mismatched heteroduplex (no MutS protein) should also be run. Incubate all reactions on ice for 30 min.

To 3 μl of the DNA/MutS mixture from each incubation add 1 μl of a 50% w/v sucrose solution.

Load 2 μl of the mixture from Step 2 onto a 6% non-denaturing polyacrylamide gel prepared in Tris-acetate-EDTA (TAE) buffer (Sambrook et al., "Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor Laboratory, New York (1989)) to which MgCl_2 has been added to a final concentration of 1 mM and run the gel at 10 V/cm and 4° C. in TAE buffer containing 1 mM MgCl_2 until bromophenol blue dye (loaded into an adjacent well) has migrated approximately half way down the gel. The presence of Mg^{++} in the gel and running buffer is critical for optimal results in the GSMA assay of MutS protein.

Filter Binding Assay

The total binding reaction volume is 10 μl . It consists of 6 μl , or 1.2 pmols, of duplex DNA and 4 μl of a MutS protein dilution containing 0.5–5 pmols (0.125–1.25 units) of MutS protein (1 pmol=97 ng). Each type of duplex, complementary and mismatched, should be assayed in duplicate or triplicate along with a no protein control for

40

each annealing, which will serve as the background to subtract.

In order to use the filter binding assay it will be necessary to make up additional annealing buffer for use in the washing step. Add 20 ml of 1M Tris-HCl, pH 7.6, 5 ml of 1M MgCl_2 , 0.1 ml of 1M DTT, and 0.02 ml of 0.5M EDTA to distilled water and bring the volume to 1 liter.

For each binding assay, add the following to a 0.5 ml microcentrifuge tube on ice:

MUTS1/MUTS2 (Control) OR

MUTS1/MUTS3 (Mismatched)

Annealing Mixture 6 μl

Set up the filtration apparatus and presoak the nitrocellulose filters in annealing buffer.

Add 4 μl of MutS protein dilution to the annealing mixtures on ice. Also include no protein controls for each annealing.

After 30 minutes, begin filtration of samples. Caution, use a slow rate of filtration. It should take at least a second or two for the 10 μl sample to filter.

Immediately wash the filters with 5 ml each of cold annealing buffer. This should take 20–30 seconds.

Place the filters in liquid scintillation vials, add fluid and count for 2 minutes each.

Determine the input cpm for each annealing as follows: To 6 μl of annealing mixture, add 54 μl of water and count 2–3 aliquots of 6 μl each in scintillation fluid. The input cpm is then 10 \times the average of the cpm of the dilution.

Determine the cpm/pmol of DNA as follows:

$$\frac{\text{cpm of 6 } \mu\text{l aliquot} \times \text{dilution} \times \text{fraction of label incorporate}}{\text{pmol of DNA in annealing reaction}}$$

A 6 μl annealing contains 1.2 pmols of DNA

A typical kinase reaction may give 42% incorporation (determined previously)

A 6 μl aliquot of 10 \times dilution may be 10,600 cpm

$$\frac{10,600 \times 10 \times 0.42}{1.2} = 37,100 \text{ cpm/pmol DNA}$$

Determine the pmols of DNA bound by various pmols of MutS. First, determine the pmols of MutS protein in a binding reaction:

$$\frac{\text{concentration of MutS} \times \text{volume of protein added}}{\text{molecular weight of MutS} \times \text{dilution factor}}$$

Example: If 4 μl of a 6 \times dilution of MutS at 250 $\mu\text{g/ml}$ is used, then:

5,556,750

41

$$\frac{250 \text{ ng/}\mu\text{l} \times 4 \mu\text{l}}{97 \text{ ng/pmol} \times 6} = 1.72 \text{ pmoles of MutS in reaction}$$

Then, determine the pmoles of DNA bound:

$$\frac{\text{cpm retained on filter with MutS}}{\text{protein - cpm on no protein filter}} = \frac{\text{cpm/pmol of DNA}}{\text{cpm/pmol of DNA}}$$

Example: One gets 15,470 cpm on the filter with MutS and 340 cpm with no protein

$$\frac{15,470 \text{ cpm} - 340 \text{ cpm}}{37,100 \text{ cpm/pmol}} = 0.408 \text{ pmoles of DNA bound}$$

Determine the number of pmoles of MutS required to bind 1 pmole of DNA (i.e., a unit of MutS).
In the above example, 1.72 pmoles of MutS bound 0.408 pmoles of DNA, such that one unit = $1.72/0.408 = 4.2$ pmoles MutS per mole DNA.

EXAMPLE 5

Effects of MutS and MutL on RecA-catalyzed Strand Transfer

A model system used to evaluate MutS and MutL effects on RecA catalyzed strand transfer is depicted in FIG. 6. The assay for RecA-catalyzed strand transfer between homologous and quasi-homologous DNA sequences employed the three strand reaction in which one strand from a linear duplex DNA is transferred to an homologous, single-stranded DNA circle (Cox, 78 *Proc. Natl. Acad. Sci. USA* 3433, 1981. These experiments exploited the previous observation that RecA is able to support strand transfer between related fd and M13 DNAs (Bianchi et al., 35 *Cell* 511, 1983; DasGupta et al., 79 *Proc. Natl. Acad. Sci. USA* 762, 1982, which are approximately 97% homologous at the nucleotide level. The vast majority of this variation is due to single base pair changes.

Results of experiments on the effects of MutS and MutL on RecA-catalyzed strand transfer between homologous and quasi-homologous DNA sequences are shown in FIG. 7. Reactions (50 μl) contained 50 mM HEPES (pH 7.5), 12 mM MgCl_2 , 2 mM ATP, 0.4 mM dithiothreitol, 6 mM phosphocreatine, 10 U/ml phosphocreatine kinase, 0.6 nM single-stranded circular DNA (molecules), 7.6 μg RecA protein, 0.54 μg SSB, and MutS or MutL as indicated. Reactions were allowed to preincubate at 37° C. for 10 minutes, strand exchange was initiated by addition of linear duplex fd DNA (Rf DNA linearized by cleavage with HpaI, 0.6 nM final concentration as molecules), and incubation continued for 70 minutes. MutS or MutL was added 1 minute prior to addition of duplex DNA. Sample (50 μl) were quenched by addition of EDTA (25 mM), sodium dodecyl sulphate (0.1%), and proteinase K (150 $\mu\text{g/ml}$), followed by incubations at 42° C. for 30 minutes.

The presence of MutS or MutL was without significant effect on strand transfer between linear duplex fd DNA and circular fd single-strands, MutS did inhibit strand transfer between quasi-homologous linear duplex fd DNA and M13 single-strands. Similar results were obtained for strand transfer between duplex M13 DNA and single-stranded fd (data not shown). In contrast, MutL alone did not significantly alter the yield of circular duplex product formed by RecA catalyzed strand transfer between these different DNAs.

42

EXAMPLE 6

MutL Potentiation of MutS Block to Strand Transfer

Results of experiments on the MutL potentiation of the MutS block to strand transfer in response to mismatched base pairs are shown in FIG. 8. Reaction mixtures (210 μl) contained 50 mM HEPES (pH 7.5), 12 mM MgCl_2 , 2 mM ATP, 0.4 mM dithiothreitol, 6 mM phosphocreatine, 10 U/ml phosphocreatine kinase, 0.6 nM (molecules) single-stranded circular DNA, 32 μg recA protein, and 2.3 μg SSB. Reactions were preincubated for 10 minutes at 37° C. and strand exchange initiated by addition of duplex fd DNA (Rf DNA linearized by cleavage with HpaI, 0.6 nM final concentration as molecules). When present, MutS (2.9 μg) and/or MutL (1.3 μg) were added 1 minute prior to addition of duplex DNA. Samples were removed as indicated times and quenched as described in Example 5.

MutL potentiates the inhibition of heteroduplex formation that is observed with MutS. Formation of full length, circular heteroduplex product is virtually abolished in the presence of MutS and MutL. Heteroduplex formation between perfectly homologous strands occurred readily in the presence of either or both proteins.

EXAMPLE 7

MutS and MutL Block of Branch Migration

While MutS and MutS along with MutL blocked formation of fully duplex, circular fd-M13 product, some strand transfer did occur in these reactions as demonstrated by the occurrence of strand transfer "intermediates" that migrated more slowly in agarose gels than fully duplex, nicked circular product (data not shown). The nature of these structures was examined using the S1 nuclease procedure of Cox and Lehman to evaluate mean length of stable heteroduplex formation. This analysis is shown in FIG. 9.

Reaction mixtures (510 μl) contained 50 mM HEPES (pH 7.5), 12 mM MgCl_2 , 2 mM ATP, 0.4 mM dithiothreitol, 6 mM phosphocreatine, 10 U/ml phosphocreatine kinase, 0.6 nM single-stranded circular DNA (molecules), 77 μg RecA protein, 5.5 μg SSB, and when indicated 6.9 μg MutS and 3.2 μg MutL. Reactions were allowed to preincubate at 37° C. for 10 minutes, strand exchange was initiated by addition of linear duplex [^3H]M13 DNA (Rf DNA linearized by cleavage with HpaI, 0.6 nM final concentration as molecules). MutS or MutL was added 1 minute prior to addition of M13 duplex DNA. Samples (100 μl) were taken as indicated, quenched with sodium dodecyl sulphate (0.8%), and extracted with phenol:chloroform:isoamyl alcohol (24:24:1) equilibrated with 10 mM Tris-HCl, pH 8.0, 0.1 mM EDTA. The organic phase was back-extracted with 0.5 volume of 50 mM HEPES, pH 5.5. Aqueous layers were combined washed with H₂O-saturated ether, and relieved of residual ether by 30 minutes incubation at 37° C. The mean length of stable heteroduplex was then determined using S1 nuclease (10 U/ml) according to Cox and Lehman (Cox, 1981 *supra*).

Although some strand transfer occurs between fd and M13 DNAs in the presence of MutS and MutL, heteroduplex formation is restricted to about one kilobase of the 6.4 kilobase possible. The MutS and MutL effects on recombination are due, at least in part, to their ability to control branch migration reaction in response to occurrence of mismatched base pairs.

Other embodiments are within the following claims.

5,556,750

43

What is claimed is:

1. A method for eliminating DNA molecules containing one or more base pairing mismatches from a population of heterohybrid duplex DNA molecules formed by base pairing of single-stranded DNA molecules obtained from a first source and a second source, comprising the steps of:
 - digesting genomic DNA from said first and said second source with a restriction endonuclease,
 - methylating the DNA from one of said sources,
 - denaturing said DNA from said first and said second source,
 - mixing DNA from said first and said second source in the presence of a recombinase protein, proteins of a mismatch repair system that modulate said recombinase protein, single-strand binding protein, and ATP, such that DNA duplexes form in homologous regions of DNA molecules from said first and said second source and the presence of a base pair mismatch results in regions that remain single-stranded, and
 - removing molecules that contain a said single-stranded region from said population.
2. The method of claim 1, wherein said recombinase protein is the *Escherichia coli* RecA protein.
3. The method of claim 1, wherein said mismatch repair system comprises the *Escherichia coli* methyl-directed mismatch repair system and consists of the MutS and MutL proteins.
4. A method for eliminating DNA molecules containing one or more mismatches from a population of heterohybrid duplex DNA molecules formed by a strand transfer reaction between duplex DNA molecules obtained from a first source and denatured DNA molecules obtained from a second source, comprising the steps of:
 - restriction digesting genomic DNA from said first and said second sources,
 - methylating the DNA of said first or said second source, denaturing DNA from said second source,
 - mixing DNA molecules from said first and said second source in the presence of a protein which catalyzes a strand transfer reaction, proteins of a mismatch repair system that modulate said protein which catalyzes a strand transfer reaction, single-strand binding protein, and ATP, such that DNA heteroduplexes form in homologous regions of DNA molecules from the first and the second source and the presence of a base pair mismatch results in regions that remain single-stranded, and
 - removing molecules that contain a said single-stranded region from said population.
5. The method of claim 1 or 4, wherein the removal of said molecules containing a single-stranded region is by treatment with a single-strand specific nuclease.
6. The method of claim 1 or 4, wherein the removal of said molecules containing a single-stranded region is by chromatography on benzoyleated naphthoylated DEAE.
7. The method of claim 4, wherein said strand transferase protein is the *Escherichia coli* RecA protein.
8. The method of claim 4, wherein mismatch repair system comprises the *Escherichia coli* methyl-directed mismatch repair system and consists of the MutS and MutL proteins.
9. The method of claim 1 or 4, wherein said sources of DNA are selected from the group consisting of: individuals of the same species, individuals of different species, indi-

44

viduals of different kingdoms, different tissue types, the same tissue type in different states of growth, different cell types, cells of the same type in different states of growth, cells of the same origin in different stages of development and cells of the same origin have undergone differential somatic mutagenesis.

10. The method of claim 9, wherein one of said sources consists of a probe sequence that has been at least partially characterized.

11. A method of genomic mismatch scanning, wherein heterohybrid DNA molecules containing base pair mismatches are removed, without the use of exonuclease III, comprising the steps of:

contacting a population of heterohybrid DNA molecules potentially containing a base pair mismatch with a DNA mismatch repair system in the presence of one or more dideoxynucleoside triphosphates such that a single-stranded region is generated in a DNA molecule that contained a base pair mismatch, without the use of exonuclease III, and,

removing said molecule containing a single-stranded region from the population.

12. The method of claim 11, wherein said DNA mismatch repair system is the *Escherichia coli* methyl-directed mismatch repair system.

13. The method of claim 11, wherein the removal of said molecule containing a single-stranded region is by chromatography on benzoyleated naphthoylated DEAE.

14. The method of claim 11, wherein the removal of said molecule containing a single-stranded region is by treatment with a single-strand specific nuclease.

15. A method for fractionating a population of DNA molecules based upon a mismatch in a subset of said molecules, wherein a heterohybrid DNA molecule containing a base pair mismatch is separated from non-mismatch-containing molecules, without the use of exonuclease III, comprising the steps of:

contacting a population of heterohybrid DNA molecules potentially containing a base pair mismatch with a DNA mismatch repair system and biotinylated nucleotide triphosphates, in the absence of exonuclease III, such that biotinylated nucleotides are incorporated into DNA molecules that contained a base pair mismatch and,

separating said molecule containing biotinylated nucleotides from those not containing said nucleotides by binding to avidin.

16. The method of claim 15, wherein said DNA mismatch repair system is the *Escherichia coli* methyl-directed mismatch repair system.

17. A method for removing DNA duplex molecules containing base pair mismatches in a population of heteroduplex DNA molecules produced from different sources, comprising the steps of:

contacting said population of DNA duplex molecules potentially containing base pair mismatches with some or all components of a mismatch repair system and biotinylated nucleotides such that said nucleotides are incorporated into duplex molecules that contain a base pair mismatch and not into DNA duplex molecules lacking a base pair mismatch, and

removing said duplexes with incorporated biotinylated nucleotides by binding to avidin.

* * * * *

EXHIBIT C



US005679522A

United States Patent [19][11] **Patent Number:** **5,679,522****Modrich et al.**[45] **Date of Patent:** ***Oct. 21, 1997**

[54] **METHODS OF ANALYSIS AND
MANIPULATION OF DNA UTILIZING
MISMATCH REPAIR SYSTEMS**

[75] **Inventors:** Paul L. Modrich, Chapel Hill, N.C.;
Shin-San Su, Newton, Mass.; Karin G.
Au, Durham, N.C.; Robert S. Lahue,
Northboro; Deani Lee Cooper,
Watertown, both of Mass.; Leroy
Worth, Jr., Durham, N.C.

[73] **Assignee:** Duke University, Durham, N.C.

[*] **Notice:** The term of this patent shall not extend
beyond the expiration date of Pat. No.
5,459,039.

[21] **Appl. No.:** 459,409

[22] **Filed:** Jun. 2, 1995

Related U.S. Application Data

[60] Division of Ser. No. 145,837, Nov. 1, 1993, which is a
continuation-in-part of Ser. No. 2,529, Jan. 11, 1993, aban-
doned, which is a continuation of Ser. No. 350,983, May 12,
1989, abandoned.

[51] **Int. Cl.⁶** C12Q 1/68; C12Q 1/70;
C12P 19/34; C07H 21/04

[52] **U.S. Cl.** 435/6; 435/91.1; 435/91.2;
435/174; 536/22.1; 536/23.1; 536/24.3;
536/24.32; 536/24.33; 530/300

[58] **Field of Search** 435/6, 91.1, 91.2,
435/174, 7.1; 536/22.1, 23.1, 24.3-33;
530/300

[56] References Cited

U.S. PATENT DOCUMENTS

4,794,075	12/1988	Ford et al.	435/6
4,818,685	4/1989	Sirover et al.	435/7
5,296,231	3/1994	Yarosh et al.	424/450
5,459,039	10/1995	Modrich et al.	435/6

FOREIGN PATENT DOCUMENTS

2239456	7/1991	United Kingdom .
9302216	2/1993	WIPO .
9320233	10/1993	WIPO .
9322457	11/1993	WIPO .
9322462	11/1993	WIPO .

OTHER PUBLICATIONS

Jiricny et al. NAR 16: 7843-785 1988.
Revzin et al. Biotechniques 7:346 1989.
Pang et al., J. of Bacteriol. 163: 1007-1015 1985.
Preibe et al., J. of Bacteriol 170: 190-196, 1988.
Chen et al. Science 237: 1197, 1987.
Myles et al. Chemical Research in Toxicology 2: 197-226
1989.
Su et al. JBC 263: 6829-6835 1988.
Lahue et al. Science 245: 160-169 1989.
Lu et al. Genomics 14: 249-255 1992.
Au et al. JBC 267: 12142-12148 1992.
Lu et al. Cell 54: 805-812 1988.
Wu et al. J. of Bacteriol. 173: 1902-1910 1991.
Modrich, *Molecular Mechanisms of DNA - Protein Inter-
action*, 1986, NIH Grant, Abstract (Source: Crisp).

(List continued on next page.)

Primary Examiner—W. Gary Jones

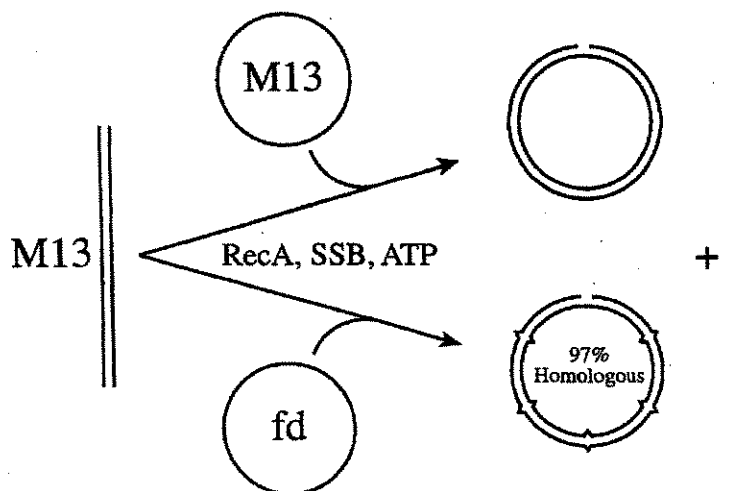
Assistant Examiner—Dianne Rees

Attorney, Agent, or Firm—Lyon & Lyon LLP

[57] ABSTRACT

A diagnostic method for detecting a base pair mismatch in a DNA duplex, comprising the steps of contacting at least one strand of a first DNA molecule with the complementary strand of a second DNA molecule under conditions such that base pairing occurs contacting a DNA duplex potentially containing a base pair mismatch with a mispair recognition protein under conditions suitable for the protein to form a specific complex only with the DNA duplex having a base pair mismatch, and not with a DNA duplex lacking a base pair mismatch, and detecting any complex as a measure of the presence of a base pair mismatch in the DNA duplex.

25 Claims, 7 Drawing Sheets



5,679,522

Page 2

OTHER PUBLICATIONS

- Myers et al., "Detection of Single Based Substitutions by Ribonuclease Cleavage at Mismatches in RNA:DNA Duplexes," *Science* 230:1242-1246 (1985).
- Nelson et al., "Genomic Mismatch Scanning A New Approach to Genetic Linkage Mapping," *Nature Genetics* 4:11-19 (1993).
- Pang et al., "Identification and Characterization of the mutL and mutS Gene Products of *Salmonella typhimurium* LT2," *J. Bacteriology* 163:1007-1015 (1985).
- Priebe et al., "Nucleotide Sequence of the hexA Gene for DNA Mismatch Repair in *Streptococcus pneumoniae* and Homology of hexA to mutS of *Escherichia coli* and *Salmonella typhimurium*," *J. Bacteriology* 170:190-196 (1988).
- Quinones et al., "Expression of the *Escherichia coli* dna Q (mutD) Gene is Inducible," *Mol. Gene Genet.* 211:106-112 (1988).
- Rayssiguier et al., "The Barrier to Recombination Between *Escherichia coli* and *Salmonella typhimurium* is Disrupted in Mismatch-Repair Mutants," *Nature* 342:396-401 (1989).
- Reenan and Kolodner, "Isolation and Characterization of Two *Saccharomyces cerevisiae* Genes Encoding Homologs of the Bacterial HexA and MutS Mismatch Repair Proteins," *Genetics* 132:963-973 (1992).
- Shen and Huang, "Effect of Base Pair Mismatches on Recombination Via the RecBCD Pathway," *Mol. Gen. Genet.* 218:358-360 (1989).
- Su and Modrich, "*Escherichia coli* mutS-encoded Protein Binds to Mismatched DNA Based Pairs," *Proc. Natl. Acad. Sci. USA* 83:5057-5061 (1986).
- Su et al., "Gap Formation is Associated With Methyl-Directed Mismatch Correction Under Conditions of Restricted DNA Synthesis," *Genome* 31:104-111 (1989).
- Su et al., "Mispair Specificity of Methyl-directed DNA Mismatch Correction in Vitro," *J. Biol. Chem.* 263:6829-6835 (1988).
- Wilchele et al., *Analytical Biochem.* 171:1-32 (1988).
- Welsh et al., "Isolation and Characterization of the *Escherichia coli* mutH Gene Product," *J. Biol. Chem.* 262:15624-15629 (1987).
- Adams et al., "The Biochemistry of the Nucleic Acids," Chapman & Hall pp. 221-223 (1986).
- Au et al., "*Escherichia coli* mutY Gene Encodes An Adenine Glycosylase Active on G-A Mispairs," *Proc. Natl. Acad. Sci. USA* 86:8877-8881 (1989).
- Au et al., "*Escherichia coli* mutY Gene Product is Required for Specific A-G C-G Mismatch Correction," *Proc. Natl. Acad. Sci. USA* 85:9163-9166 (1988).
- Au et al., "Initiation of Methyl-directed Mismatch Repair," *J. Biol. Chem.* 267:12142-12148 (1992).
- Bianchi and Radding, "Insertions, Deletions and Mismatches in Heteroduplex DNA Made by RecA Protein," *Cell* 35:511-520 (1983).
- Chen and Sigman, "Chemical Conversion of a DNA-Binding Protein Into a Site-Specific Nuclease," *Science* 237:1197-1201 (1987).
- Cooper et al., "Methyl-Directed Mismatch Repair is Bidirectional," *J. Biol. Chem.* 268:11823-11829 (1993).
- Cotton et al., "Reactivity of Cytosine and Thymine In Single-Base-Pair Mismatches with Hydroxylamine and Osmium Tetroxide and Its Application to the Study of Mutations," *Proc. Natl. Acad. Sci. USA* 85:4397-4401 (1988).
- Dasgupta and Radding, "Polar Branch Migration Promoted by recA Protein: Effect of Mismatched Base Pairs," *Proc. Natl. Acad. Sci. USA* 79:762-766 (1982).
- Fang and Modrich, "Human Strand-Specific Mismatch Repair Occurs by a Bidirectional Mechanism Similar to That of the Bacterial Reaction," *J. Biol. Chem.* 268:11838-11844 (1993).
- Fujii and Shimada, "Isolation and Characterization of cDNA Clones Derived from the Divergently Transcribed Gene in the Region Upstream from the Human Dihydrofolate Reductase Gene," *J. Biol. Chem.* 264:10057-10064 (1989).
- Grilley et al., "Isolation and Characterization of the *Escherichia coli* mutL Gene Product," *J. Biol. Chem.* 264:1000-1004 (1989).
- Grilley et al., "Mechanisms of DNA-Mismatch Correction," *Mutation Research* 236:253-267 (1990).
- Grilley et al., "Bidirectional Excision in Methyl-Directed Mismatch Repair," *J. Biol. Chem.* 268:11830-11837 (1993).
- Hennighausen and Lubon, "Interaction of Protein With DNA In Vitro," *Guide to Molecular Cloning Techniques*, Berger and Kimmel eds., 152:721-735 (1987).
- Holmes et al., "Strand-specific Mismatch Correction In Nuclear Extracts of Human and *Drosophila Melanogaster* Cell Lines," *Proc. Natl. Acad. Sci. USA* 87:5837-5841 (1990).
- Jiricny et al., "Mismatch-containing Oligonucleotide Duplexes Bound By the *E. coli* mutS-encoded Protein," *Nucleic Acids Research* 16:7843-7853 (1988).
- Lahue et al., "Requirement for d(GATC) Sequences in *Escherichia coli* mutHLS Mismatch Correction," *Proc. Natl. Acad. Sci. USA* 84:1482-1486 (1987).
- Lahue and Modrich, "DNA Mismatch Correction in a Defined System," *Science* 245:160-164 (1989).
- Lahue and Modrich, "Methyl-directed DNA Mismatch Repair in *Escherichia coli*," *Mutation Research* 198:37-43 (1988).
- Lu et al., "Repair of DNA Base-pair Mismatches in Extracts of *Escherichia coli*," *Cold Spring Harbor Laboratory, Cold Spring Harbor Symposia on Quantitative Biology*, XLIX:589-596 (1984).
- Lu and Chang, "Repair of Single Base-Pair Transversion Mismatches of *Escherichia coli* in Vitro: Correction of Certain A/G Mismatches is Independent of dam Methylation and Host mutHLS Gene Functions," *Genetics* 118:593-600 (1988).
- Lu and Chang, "A Novel Nucleotide Excision Repair for The Conversion of An A/G Mismatch to C/G Base Pair in *E. coli*," *Cell* 54:805-812 (1988).
- Lu, "Influence of GATC Sequences on *Escherichia coli* DNA Mismatch Repair In Vitro," *J. Bacteriology* 169:1254-1259 (1987).
- Lu and Hsu, "Detection of Single DNA Base Mutations with Mismatch Repair Enzymes," *Genomics* 14:249-255 (1992).
- Lu et al., "Methyl-directed Repair of DNA Base-pair Mismatches in Vitro," *Proc. Natl. Acad. Sci. USA* 80:4639-4643 (1983).
- Marx, "DNA Repair Comes Into Its Own," *Science* 266:728-730 (1994).
- Modrich, "Mechanisms and Biological Effects on Mismatch Repair," *Ann. Rev. Genet.* 25:229-253 (1991).
- Modrich et al., "DNA Mismatch Correction," *Ann. Rev. Biochem.* 56:435-466 (1987).
- Modrich, "Mismatch, Repair, Genetic Stability and Cancer," *Science* 266:1959-1960 (1994).
- Modrich, "Methyl-directed DNA Mismatch Correction," *J. Biol. Chem.* 264:6597-6600 (1989).

U.S. Patent

Oct. 21, 1997

Sheet 1 of 7

5,679,522

V 5'-AAGCTTTCGAG Hind III
C 3'-TTCGAGAGCTC Xho I

FIG. 1.

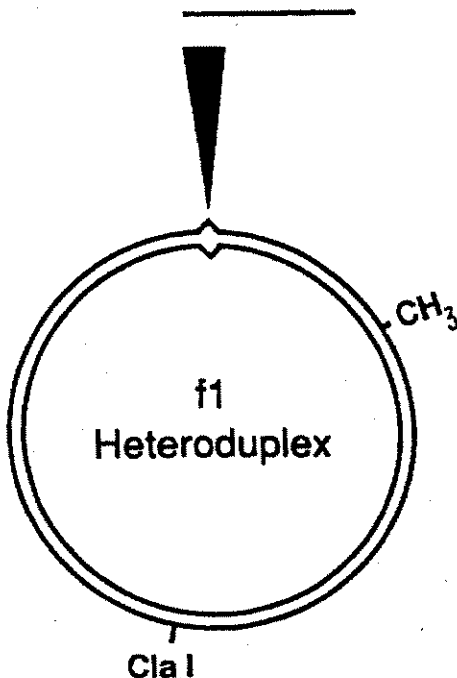
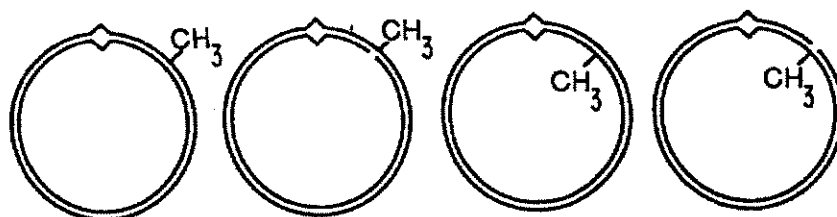


FIG. 4.



Reaction conditions	Repair (fmol/20 min)			
Complete	15 (<1)	17 (<1)	8 (<1)	10 (<1)
- Mut H	<1	18	1	9
- Mut L	<1	<1	<1	<1
- Mut S	<1	<1	<1	1
- SSB	2	<1	<1	<1
- pol III holoenzyme	<1	<1	<1	<1

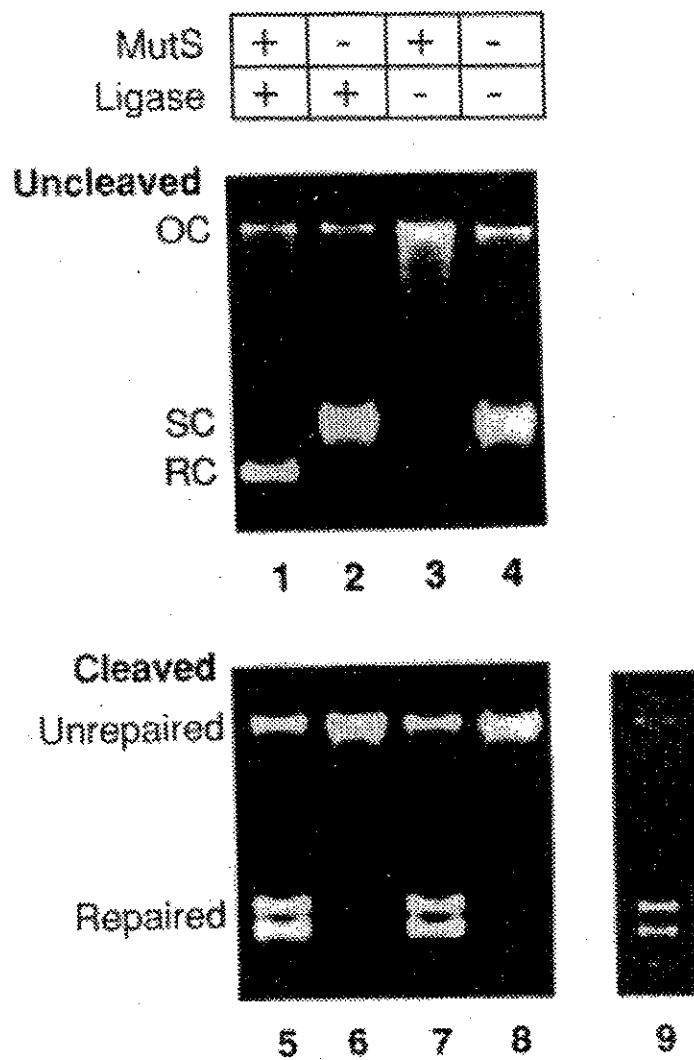
U.S. Patent

Oct. 21, 1997

Sheet 2 of 7

5,679,522

FIG. 2.



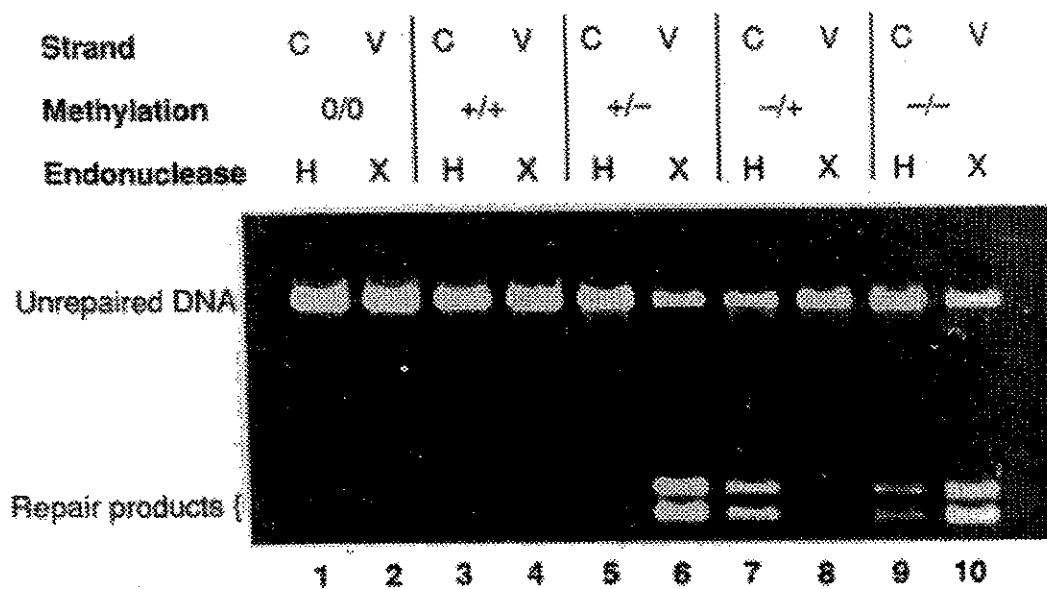
U.S. Patent

Oct. 21, 1997

Sheet 3 of 7

5,679,522

FIG. 3.

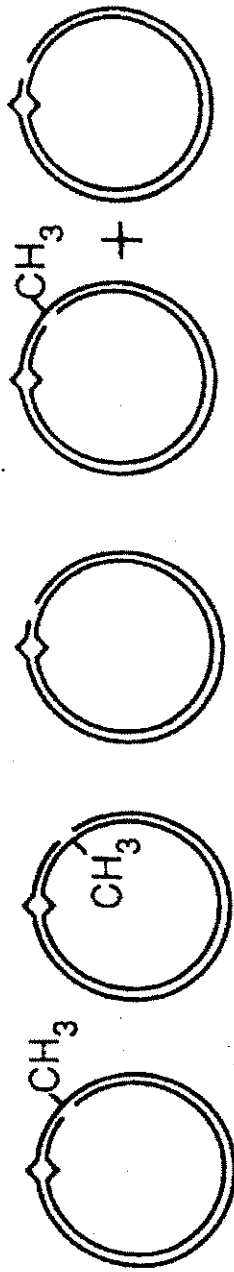


U.S. Patent

Oct. 21, 1997

Sheet 4 of 7

5,679,522



Repair (fmol/20 min)

Ligase Muth

Ligase Muth	Repair (fmol/20 min)			
	19 (<1)	9 (<1)	11 (<1)	19 (<1)
—	—	—	—	9 (<1)
+	2	<1	1	2
+	20	7	2	15

FIG. 5.

U.S. Patent

Oct. 21, 1997

Sheet 5 of 7

5,679,522

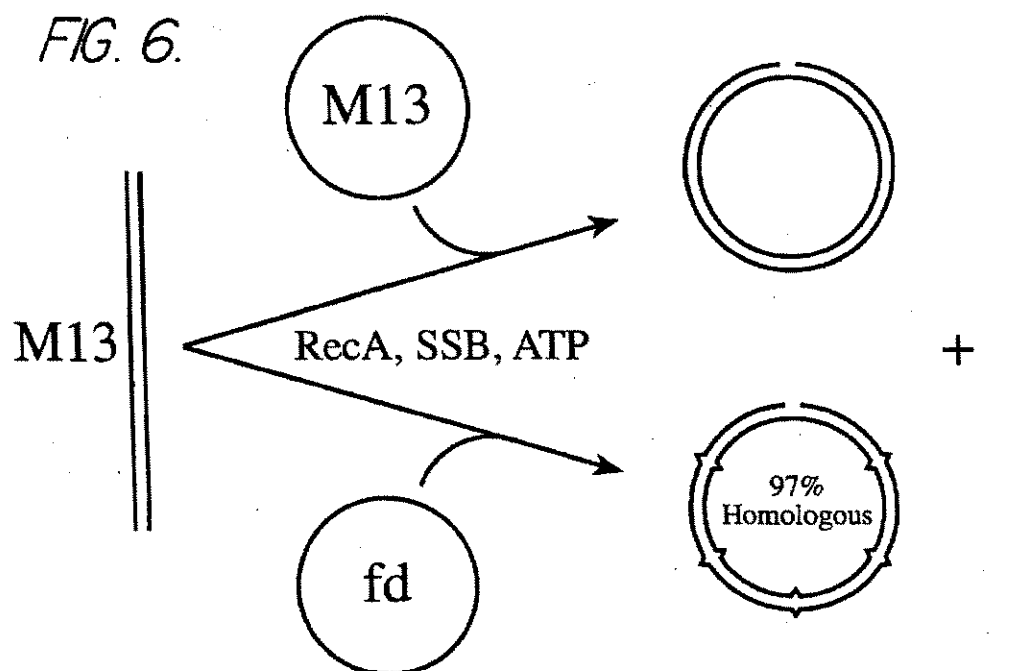
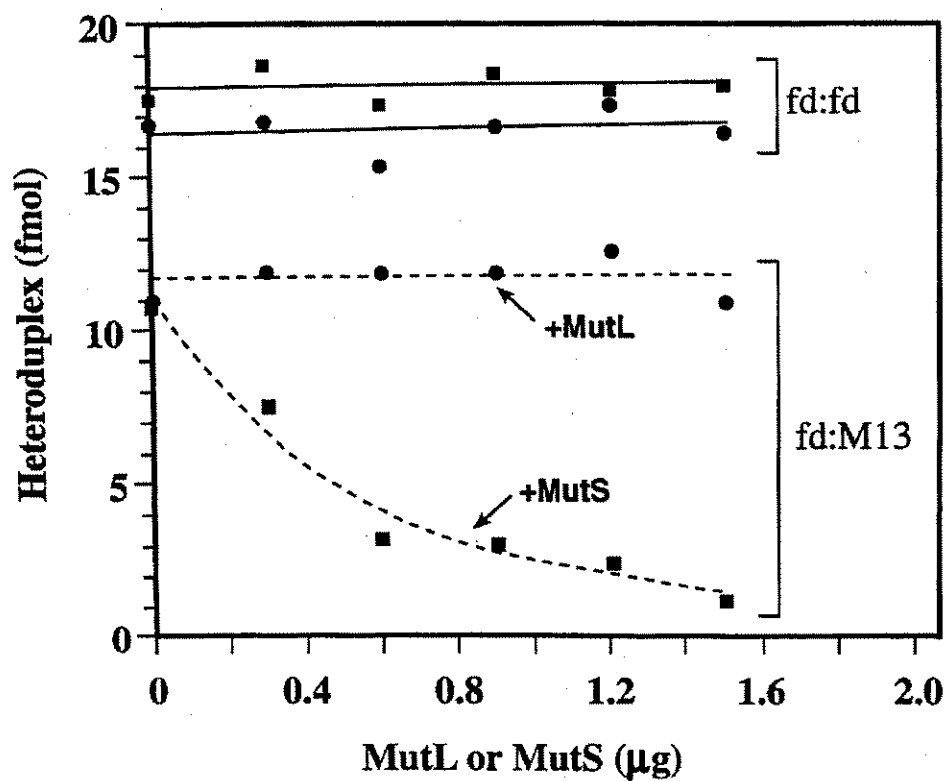


FIG. 7.



U.S. Patent

Oct. 21, 1997

Sheet 6 of 7

5,679,522

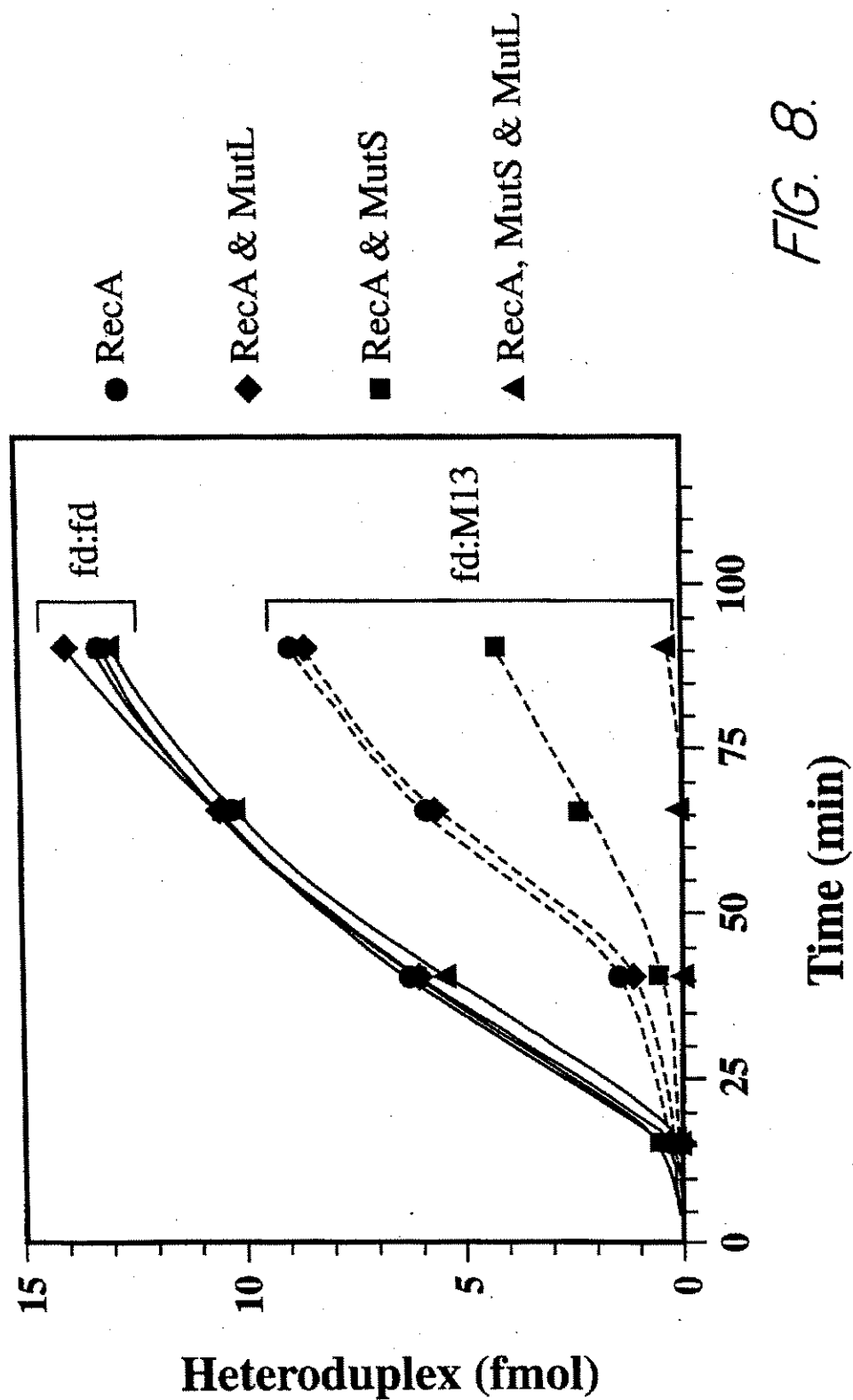


FIG. 8.

U.S. Patent

Oct. 21, 1997

Sheet 7 of 7

5,679,522

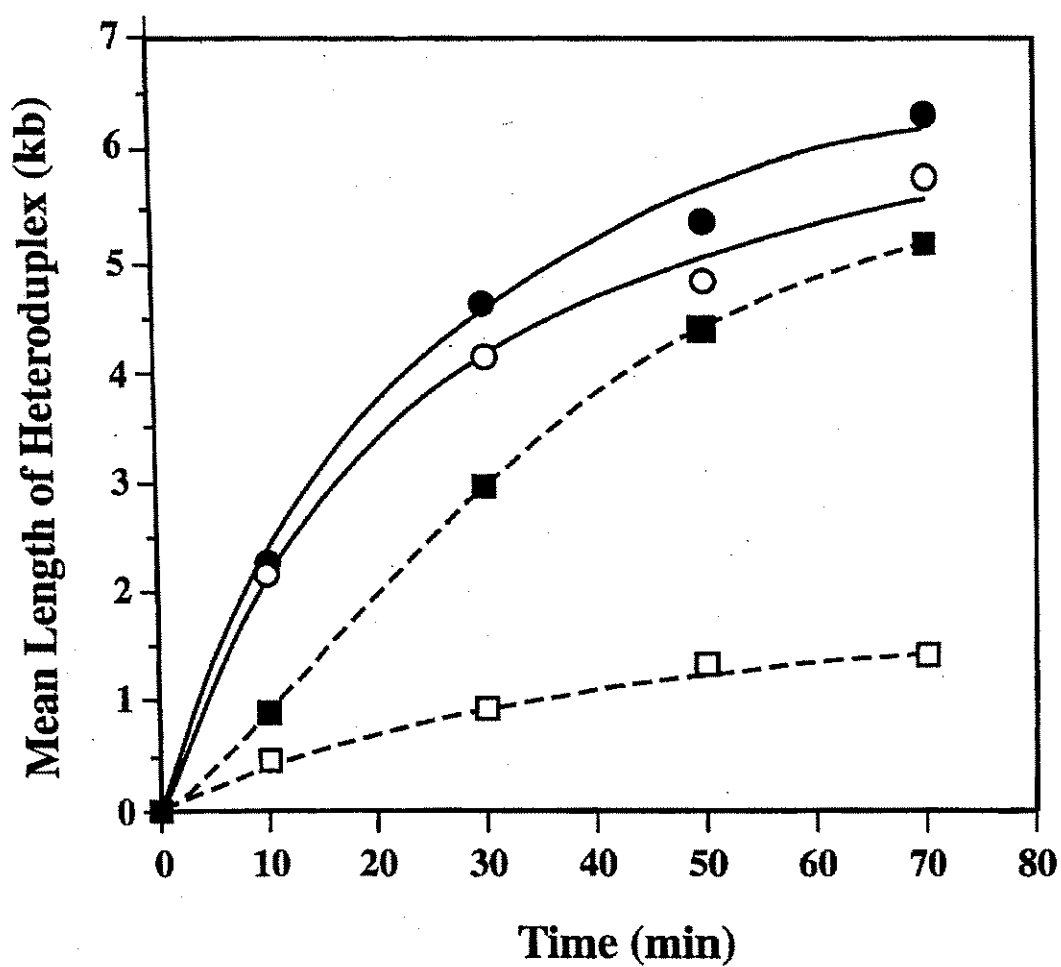


FIG. 9.

5,679,522

1

METHODS OF ANALYSIS AND MANIPULATION OF DNA UTILIZING MISMATCH REPAIR SYSTEMS

This application is a divisional of U.S. Ser. No. 08/145, 837 filed Nov. 1, 1993 which is a continuation-in-part of Modrich et al., U.S. Ser. No. 08/002,529, filed Jan. 11, 1993, now abandoned, entitled "Methods For Mapping Genetic Mutations" which is a continuation of U.S. Ser. No. 07/350, 983, filed May 12, 1989, now abandoned entitled, "Methods For Mapping Genetic Mutations", both hereby incorporated by reference herein, including drawings.

This work was supported by the U.S. government, namely Grant No. GM23719. The U.S. government may have rights in this invention.

FIELD OF THE INVENTION

The present invention relates to methods for mapping genetic differences among deoxyribonucleic acid ("DNA") molecules, especially mutations involving a difference in a single base between the base sequences of two homologous DNA molecules.

BACKGROUND OF THE INVENTION

The following is a discussion of relevant art, none of which is admitted to be prior art to the appended claims.

Mapping of genetic differences between individuals is of growing importance for both forensic and medical applications. For example, DNA "fingerprinting" methods are being applied for identification of perpetrators of crimes where even small amounts of blood or sperm are available for analysis. Biological parents can also be identified by comparing DNAs of a child and a suspected parent using such means. Further, a number of inherited pathological conditions may be diagnosed before onset of symptoms, even in utero, using methods for structural analyses of DNA. Finally, it is notable that a major international effort to physically map and, ultimately, to determine the sequence of bases in the DNA encoding the entire human genome is now underway and gaining momentum in both institutional and commercial settings.

DNA molecules are linear polymers of subunits called nucleotides. Each nucleotide comprises a common cyclic sugar molecule, which in DNA is linked by phosphate groups on opposite sides to the sugars of adjoining nucleotides, and one of several cyclic substituents called bases. The four bases commonly found in DNAs from natural sources are adenine, guanine, cytosine and thymine, hereinafter referred to as A, G, C and T, respectively. The linear sequence of these bases in the DNA of an individual encodes the genetic information that determines the heritable characteristics of that individual.

In double-stranded DNA, such as occurs in the chromosomes of all cellular organisms, the two DNA strands are entwined in a precise helical configuration with the bases projecting inward and so aligned as to allow interactions between bases from opposing strands. The two strands are held together in precise alignment mainly by hydrogen bonds which are permitted between bases by a complementarity of structures of specific pairs of bases. This structural complementarity is determined by the chemical natures and locations of substituents on each of the bases. Thus, in double-stranded DNA, normally each A on one strand pairs with a T from the opposing strand, and, likewise, each G with an opposing C.

When a cell undergoes reproduction, its DNA molecules are replicated and precise copies are passed on to its descen-

2

dants. The linear base sequence of a DNA molecule is maintained in the progeny during replication in the first instance by the complementary base pairings which allow each strand of the DNA duplex to serve as a template to align free nucleotides with its polymerized nucleotides. The complementary nucleotides so aligned are biochemically polymerized into a new DNA strand with a base sequence that is entirely complementary to that of the template strand.

Occasionally, an incorrect base pairing does occur during replication, which, after further replication of the new strand, results in a double-stranded DNA offspring with a sequence containing a heritable single base difference from that of the parent DNA molecule. Such heritable changes are called genetic mutations, or more particularly in the present case, "single base pair" or "point" mutations. The consequences of a point mutation may range from negligible to lethal, depending on the location and effect of the sequence change in relation to the genetic information encoded by the DNA.

The bases A and G are of a class of compounds called purines, while T and C are pyrimidines. Whereas the normal base pairings in DNA (A with T, G with C) involve one purine and one pyrimidine, the most common single base mutations involve substitution of one purine or pyrimidine for the other (e.g., A for G or C for T or vice versa), a type of mutation referred to as a "transition". Mutations in which a purine is substituted for a pyrimidine, or vice versa, are less frequently occurring and are called "transversions". Still less common are point mutations comprising the addition or loss of a small number (1, 2 or 3) of nucleotides arising in one strand of a DNA duplex at some stage of the replication process. Such mutations are called small "insertions" or "deletions", respectively, and are also known as "frameshift" mutations in the case of insertion/deletion of one of two nucleotides, due to their effects on translation of the genetic code into proteins. Mutations involving larger sequence rearrangement also do occur and can be important in medical genetics, but their occurrences are relatively rare compared to the classes summarized above.

Mapping of genetic mutations involves both the detection of sequence differences between DNA molecules comprising substantially identical (i.e., homologous) base sequences, and also the physical localization of those differences within some subset of the sequences in the molecules being compared. In principle, it is possible to both detect and localize limited genetic differences, including point mutations within genetic sequences of two individuals, by directly comparing the sequences of the bases in their DNA molecules.

Other methods for detecting differences between DNA sequences have been developed. For example, some pairs of single-stranded DNA fragments with sequences differing in a single base may be distinguished by their different migration rates in electric fields, as in denaturing gradient gel electrophoresis.

DNA restriction systems found in bacteria for example, comprise proteins which generally recognize specific sequences in double-stranded DNA composed of 4 to 6 or more base pairs. In the absence of certain modifications (e.g., a covalently attached methyl group) at definite positions within the restriction recognition sequence, endonuclease components of the restriction system will cleave both strands of a DNA molecule at specific sites within or near the recognition sequence. Such short recognition sequences occur by chance in all natural DNA sequences, once in every few hundred or thousand base pairs, depending on the

5,679,522

3

recognition sequence length. Thus, digestion of a DNA molecule with various restriction endonucleases, followed by analyses of the sizes of the resulting fragments (e.g., by gel electrophoresis), may be used to generate a physical map ("fingerprint") of the locations in a DNA molecule of selected short sequences.

Comparisons of such restriction maps of two homologous DNA sequences can reveal differences within those specific sequences that are recognized by those restriction enzymes used in the available maps. Restriction map comparisons may localize any detectable differences within limits defined ultimately by the resolving power of DNA fragment size determination, essentially within about the length of the restriction recognition sequence under certain conditions of gel electrophoresis.

In practice, selected heritable differences in restriction fragment lengths (i.e., restriction fragment length polymorphisms, "RFLP"s) have been extremely useful, for instance, for generating physical maps of the human genome on which genetic defects may be located with a relatively low precision of hundreds or, sometimes, tens of thousands of base pairs. Typically, RFLPs are detected in human DNA isolated from small tissue or blood samples by using radioactively labeled DNA fragments complementary to the genes of interest. These "probes" are allowed to form DNA duplexes with restriction fragments of the human DNA after separation by electrophoresis, and the resulting radioactive duplex fragments are visualized by exposure to photographic (e.g., X-ray sensitive) film, thereby allowing selective detection of only the relevant gene sequences amid the myriad of others in the genomic DNA.

When the search for DNA sequence differences can be confined to specific regions of known sequence, the recently developed "polymerase chain reaction" ("PCR") technology can be used. Briefly, this method utilizes short DNA fragments complementary to sequences on either side of the location to be analyzed to serve as points of initiation for DNA synthesis (i.e., "primers") by purified DNA polymerase. The resulting cyclic process of DNA synthesis results in massive biochemical amplification of the sequences selected for analysis, which then may be easily detected and, if desired, further analyzed, for example, by restriction mapping or direct DNA sequencing methods. In this way, selected regions of a human gene comprising a few kbp may be amplified and examined for sequence variations.

Another known method for detecting and localizing single base differences within homologous DNA molecules involves the use of a radiolabeled RNA fragment with base sequence complementary to one of the DNAs and a nuclease that recognizes and cleaves single-stranded RNA. The structure of RNA is highly similar to DNA, except for a different sugar and the presence of uracil (U) in place of T; hence, RNA and DNA strands with complementary sequences can form helical duplexes ("DNA:RNA hybrids") similar to double-stranded DNA, with base pairing between A's and U's instead of A's and T's. It is known that the enzyme ribonuclease A ("RNase A") can recognize some single pairs of mismatched bases (i.e., "base mismatches") in DNA:RNA hybrids and can cleave the RNA strand at the mispair site. Analysis of the sizes of the products resulting from RNase A digestion allows localization of single base mismatches, potentially to the precise sequence position, within lengths of homologous sequences determined by the limits of resolution of the RNA sizing analysis (Myers, R. M. et al., 1985, Science, 230, 1242-1246). RNA sizing is performed in this method by standard gel electrophoresis procedures used in DNA sequencing.

4

S1 nuclease, an endonuclease specific for single-stranded nucleic acids, can recognize and cleave limited regions of mismatched base pairs in DNA:DNA or DNA:RNA duplexes. A mismatch of at least about 4 consecutive base pairs actually is generally required for recognition and cleavage of a duplex by S1 nuclease.

Ford et al., (U.S. Pat. No. 4,794,075) disclose a chemical modification procedure to detect and localize mispaired guanines and thymidines and to fractionate a pool of hybrid DNA from two samples obtained from related individuals. Carbodiimide is used to specifically derivatize unpaired G's and T's, which remain covalently associated with the DNA helix.

The present invention concerns use of proteins that function biologically to recognize mismatched base pairs in double-stranded DNA (and, therefore, are called "mispair recognition proteins") and their application in defined systems for detecting and mapping point mutations in DNAs. Accordingly, it is an object of the present invention to provide methods for using such mispair recognition proteins, alone or in combination with other proteins, for detecting and localizing base pair mismatches in duplex DNA molecules, particularly those DNAs comprising several kbp, and manipulating molecules containing such mismatches. Additionally, it is an object of this invention to develop modified forms of mispair recognition proteins to further simplify methods for identifying specific bases which differ between DNAs. The following is a brief outline of the art regarding mispair recognition proteins and systems, none of which is admitted to be prior art to the present invention.

Enzymatic systems capable of recognition and correction of base pairing errors within the DNA helix have been demonstrated in bacteria, fungi and mammalian cells, but the mechanisms and functions of mismatch correction are best understood in *Escherichia coli*. One of the several mismatch repair systems that have been identified in *E. coli* is the methyl-directed pathway for repair of DNA biosynthetic errors. The fidelity of DNA replication in *E. coli* is enhanced 100-1000 fold by this post-replication mismatch correction system. This system processes base pairing errors within the helix in a strand-specific manner by exploiting patterns of DNA methylation. Since DNA methylation is a post-synthetic modification, newly synthesized strands temporarily exist in an unmethylated state, with the transient absence of adenine methylation on GATC sequences directing mismatch correction to new DNA strands within the hemimethylated duplexes.

In vivo analyses in *E. coli* have shown that selected examples of each of the different mismatches are subject to correction with different efficiencies. G-T, A-C, G-G and A-A mismatches are typically subject to efficient repair. A-G, C-T, T-T and C-C are weaker substrates, but well repaired exceptions exist within this class. The sequence environment of a mismatched base pair may be an important factor in determining the efficiency of repair in vivo. The mismatch correction system is also capable in vivo of correcting differences between duplexed strands involving a single base insertion or deletion. Further, genetic analyses have demonstrated that the mismatch correction process requires intact genes for several proteins, including the products of the mutH, mutL and mutS genes, as well as DNA helicase II and single-stranded DNA binding protein (SSB). The following are further examples of art discussing this subject matter.

Lu et al., 80 *Proc. Natl. Acad. Sci. USA* 4639, 1983 disclose the use of a soluble *E. coli* system to support mismatch correction in vitro.

5,679,522

5

Pans et al., 163 *J. Bact.* 1007, 1985 disclose cloning of the mutS and mutL genes of *Salmonella typhimurium*.

The specific components of the *E. coli* mismatch correction system have been isolated and the biochemical functions determined. Preparation of MutS protein substantially free of other proteins has been reported (Su and Modrich, 1986, *Proc. Nat. Acad. Sci. U.S.A.*, 84, 5057-5061, which is hereby incorporated herein by reference). The isolated MutS protein was shown to recognize four of the eight possible mismatched base pairs (specifically, G-T, A-C, A-G and C-T mispairs).

Su et al., 263 *J. Biol. Chem.* 6829, 1988 disclose that the mutS gene product binds to each of the eight base pair mismatches and does so with differential efficiency.

Jiricny et al., 16 *Nucleic Acids Research* 7843, 1988 disclose binding of the mutS gene product of *E. coli* to synthetic DNA duplexes containing mismatches to correlate recognition of mispairs and efficiency of correction in vivo. Nitrocellulose filter binding assays and band-shift assays were utilized.

Welsh et al., 262 *J. Biol. Chem.* 15624, 1987 purified the product of the MutH gene to near homogeneity and demonstrated the MutH gene product to be responsible for d(GATC) site recognition and to possess a latent endonuclease that incises the unmethylated strand of hemimethylated DNA 5' to the G d(GATC) sequences.

Au et al., 267 *J. Biol. Chem.* 12142, 1992 indicate that activation of the Mute endonuclease requires MutS, MutL and ATP.

Grilley et al. 264 *J. Biol. Chem.* 1000, 1989 purified the *E. coli* mutL gene product to near homogeneity and indicate that the mutL gene product interacts with MutS heteroduplex DNA complex.

Lahue et al., 245 *Science* 160, 1989 delineate the components of the *E. coli* methyl-directed mismatch repair system that function in vitro to correct seven of the eight possible base pair mismatches. Such a reconstituted system consists of MutH, MutL, and MutS proteins, DNA helicase II, single-strand DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, DNA ligase, ATP, and the four deoxyribonucleoside triphosphates.

Su et al., 31 *Genome* 104, 1989 indicate that under conditions of restricted DNA synthesis, or limiting concentration of dNTPs, or by supplementing a reaction with a ddNTP, there is the formation of excision tracts consisting of single-stranded gaps in the region of the molecule containing a mismatch and a d(GATC) site.

Grilley et al. 268 *J. Biol. Chem.* 11830, 1993, indicate that excision tracts span the shorter distance between a mismatch and the d(GATC) site, indicating a bidirectional capacity of the methyl-directed system.

Holmes et al., 87 *Proc. Natl. Acad. Sci. USA*, 5837, 1990, disclose nuclear extracts derived from HeLa and *Drosophila melanogaster* K_c cell lines to support strand mismatch correction in vitro.

Cooper et al., 268 *J. Biol. Chem.*, 11823, 1993, describe a role for RecJ and Exonuclease VII as a 5' to 3' exonuclease in a mismatch repair reaction. In reconstituted systems such a 5' to 3' exonuclease function had been provided by certain preparations of DNA polymerase III holoenzyme.

Au et al., 86 *Proc. Natl. Acad. Sci. USA* 8877, 1989 describe purification of the MutY gene product of *E. coli* to near homogeneity, and state that the MutY protein is a DNA glycosylase that hydrolyzes the glycosyl bond linking a mispaired adenine (G-A) to deoxyribose. The MutY protein,

6

an apurinic endonuclease, DNA polymerase I, and DNA ligase were shown to reconstitute G-A to G-C mismatch correction in vitro.

A role for the *E. coli* mismatch repair system in controlling recombination between related but non allelic sequences has been indicated (Feinstein and Low, 113 *Genetics* 13, 1986; Rayssiguier, 342 *Nature* 396, 1989; Shen, 218 *Mol. Gen. Genetics* 358, 1989; Petit, 129 *Genetics* 327, 1991). The frequency of crossovers between sequences which differ by a few percent or more at the base pair level are rare. In bacterial mutants deficient in methyl-directed mismatch repair, the frequency of such events increases dramatically. The largest increases are observed in MutS and MutL deficient strains. (Rayssiguier, supra; and Petit, supra.)

Nelson et al., 4 *Nature Genetics* 11, 1993, disclose a genomic mismatch (GMS) method for genetic linkage analysis. The method allows DNA fragments from regions of identity-by-descent between two relatives to be isolated based on their ability to form mismatch-free hybrid molecules.

The method consists of digesting DNA from the two sources with a restriction endonuclease that produces protruding 3' ends. The protruding 3' ends provide some protection from exonuclease III, which is used in later steps. The two sources are distinguished by methylating the DNA from only one source. Molecules from both sources are denatured and reannealed, resulting in the formation of four types of duplex molecules: homohybrids formed from strands derived from the same source and heterohybrids consisting of DNA strands from different sources. Heterohybrids can either be mismatch free or contain base-pair mismatches, depending on the extent of identity of homologous regions.

Homohybrids are distinguished from heterohybrids by use of restriction endonucleases that cleave at fully methylated or unmethylated GATC sites. Homohybrids are cleaved to smaller duplex molecules, while heterohybrids are resistant to cleavage. Heterohybrids containing a mismatch (es) are distinguished from mismatch free molecules by use of the *E. coli* methyl-directed mismatch repair system. The combination of three proteins of the methyl-directed mismatch repair system MutH, MutL, and MutS along with ATP introduce a single-strand nick on the unmethylated strand at GATC sites in duplexes that contain a mismatch. Heterohybrids that do not contain a mismatch are not nicked. All molecules are then subject to digestion by Exonuclease III (Exo III), which can initiate digestion at a nick, a blunt end or a 5' overhang, to produce single-stranded gaps. Only mismatch free heterohybrids are not subject to attack by Exo III, all other molecules have single-stranded gaps introduced by the enzyme. Molecules with single-stranded regions are removed by absorption to benzoylated naphthoylated DEAE cellulose. The remaining molecules consist of mismatch-free heterohybrids which may represent regions of identity by descent.

SUMMARY OF THE INVENTION

Applicant has determined that a single DNA base mismatch recognition protein can form specific complexes with any of the eight possible mismatched base pairs embedded in an otherwise homologous DNA duplex. It has also been revealed that another mismatch recognition protein can recognize primarily one specific base pair mismatch, A-G, and in so doing, it chemically modifies a nucleotide at the site of the mispair. In addition, defined in vitro systems have been established for carrying out methyl-directed mismatch

5,679,522

7

repair processes. Accordingly, the present invention features the use of such mispair recognition proteins and related correction system components to detect and to localize point mutations in DNAs. In addition the invention concerns methods for the analysis and manipulation of populations of DNA duplex molecules potentially containing base pair mismatches through the use of all or part of defined mismatch repair systems.

The invention utilizes five basic methods for heteroduplex mapping analysis, and manipulation: (i) binding of a mismatch recognition protein, e.g., MutS to DNA molecules containing one or more mispairs; (ii) cleavage of a heteroduplex in the vicinity of a mismatch by a modified form of a mismatch recognition protein; (iii) mismatch-provoked cleavage at one or more GATC sites via a mismatch repair system dependent reaction, e.g., MuthLS; (iv) formation of a mismatch-provoked gap in heteroduplex DNA via reactions of a mismatch repair system and (v) labelling of mismatch-containing nucleotides with a nucleotide analog, e.g., a biotinylated nucleotide, using a complete mismatch repair system.

For clarity in the following discussion, it should be noted that certain distinctions exist related to the fact that some proteins that recognize DNA base mispairs are merely DNA binding proteins, while others modify the DNA as a consequence of mispair recognition. Notwithstanding the fact that in the latter situation the protein modifying the DNA may be associated with the DNA only transiently, hereinafter, whether a mispair recognition protein is capable of DNA binding only or also of modifying DNA, whenever it is said that a protein recognizes a DNA mispair, this is equivalent to saying that it "forms specific complexes with" or "binds specifically to" that DNA mispair in double-stranded DNA. In the absence of express reference to modification of DNA, reference to DNA mispair recognition does not imply consequent modification of the DNA. Further, the phrase "directs modification of DNA" includes both cases wherein a DNA mispair recognition protein has an inherent DNA modification function (e.g., a glycosylase) and cases wherein the mispair recognition protein merely forms specific complexes with mispairs, which complexes are then recognized by other proteins that modify the DNA in the vicinity of the complex.

Accordingly, the present invention features a method for detecting base pair mismatches in a DNA duplex by utilizing a mismatch recognition protein that forms specific complexes with mispairs, and detecting the resulting DNA:protein complexes by a suitable analytical method.

In addition to methods designed merely to detect base pair mismatches, this invention includes methods for both detecting and localizing base pair mismatches by utilizing components of mismatch repair system.

The present invention also features mispair recognition proteins which have been altered to provide an inherent means for modifying at least one strand of the DNA duplex in the vicinity of the bound mispair recognition protein.

The present invention also concerns systems utilizing an A—G specific mispair recognition protein, for example, the *E. coli* DNA mispair recognition protein that recognizes only A—G mispairs without any apparent requirement for hemimethylation. This protein, the product of the mutY gene, is a glycosylase which specifically removes the adenine from an A—G mispair in a DNA duplex. Accordingly, this MutY protein is useful for the specific detection of A—G mispairs according to the practice of the present invention.

8

The invention also includes the combined use of components of a mismatch repair system along with a recombinase protein. The recombinase protein functions to catalyze the formation of duplex molecules starting with single-stranded molecules obtained from different sources, by a renaturation reaction. Such a recombinase protein is also capable of catalyzing a strand transfer reaction between a single-stranded molecule from one source and double-stranded molecules obtained from a different source. In the presence of a base pair mismatch, formation of duplex regions catalyzed by such a recombinase protein is inhibited by components of a mismatch repair system, e.g., *E. coli* MutS and MutL, proteins. Modulation of recombinase activity by components of a mismatch repair system may involve inhibition of branch migration through regions that generate mismatched base pairs. The combination of a DNA mismatch repair system and a recombinase system provides a very sensitive selection step allowing for the removal of molecules containing a base pair mismatch from a population of newly formed heteroduplex molecules. This procedure provides a selection scheme that can be utilized independent of or in conjunction with the actual mismatch repair reaction.

The invention also features two improvements on the genomic mismatch scanning technique (GMS) of Nelson et al. 4 *Nature Genetics* 11, 1993, used to map regions of genetic identity between populations of DNA molecules.

One improvement provided by the invention features an additional selection step, as described above, for determining genetic variation. The genomic mismatch scanning (GMS) method includes one selection step which is carried out after hybrid formation. The present invention includes an additional step that occurs during hybrid formation, through the use of a protein with recombinase activity along with components of a mismatch repair system. The increase in sensitivity for screening for genetic variation provided by the additional selection step makes possible the use of the GMS technique with larger genomes, e.g., man.

A second improvement provided by the invention features the replacement or modification of the exonuclease III digestion step employed in the GMS method. In the GMS procedure exonuclease III is used to degrade all DNA molecules, except mismatch-free heterohybrids, to molecules containing single-stranded regions, which are subsequently removed. Heterohybrids are duplex molecules which are formed in the method from two molecules which were previously base paired with other molecules (i.e., from different sources). In the instant invention this step is replaced by a procedure that employs all or some of the components of a mismatch repair system. Exo III is a 3' to 5' exonuclease specific for double-stranded DNA, which preferably initiates at blunt or 5' protruding ends. In the GMS procedure DNA molecules are digested with restriction enzymes that produce protruding 3' ends. Although molecules containing protruding 3' ends are not preferred substrates for Exo III, such molecules can be subject to limited attack by the enzyme. Thus, even mismatch-free heterohybrids will be degraded to some extent by Exo III, and will be erroneously removed from the final population of molecules representing those of identity-by-descent. The invention employs components of a mismatch repair system along with dideoxy or biotinylated nucleotide, to avoid the use of Exo III and the potential loss of heterohybrids molecules that are mismatch-free. Homohybrids are digested in the presence of helicase II by exoVI RecJ and exo I, e.g., natural exonucleases involved in the mismatch repair reaction. The invention also features a modification of

5,679,522

9

the step utilizing Exo III, consisting of ligation of duplex DNA molecules at dilute concentrations so as to form closed circular monomer molecules, thus removing any 3' ends which may be subject to degradation by Exo III.

The invention includes the use of a mismatch repair system to detect and remove or correct base pair mismatches in a population produced by the process of enzymatic amplification of nucleic acid molecules. DNA polymerase errors that occur during a cycle of enzymatic amplification can result in the presence of mismatched base pair(s) in the population of product molecules. If such errors are perpetuated in subsequent cycles they can impair the value of the final amplified product. The fidelity of the amplification method can be enhanced by including one or more components of a mismatch repair system to either correct the mismatch base pair(s) or to eliminate from the amplified population, molecules that contain mismatch base pair(s). Elimination of molecules containing a base pair mismatch can be accomplished by binding to a protein, such as MutS, or by introduction of a nick in one strand of the duplex so that a full sized product will not be produced in a subsequent round of amplification.

The invention also features methods to remove molecules containing a base pair mismatch through the binding of the mismatch to the components of the mismatch repair system or by the binding of a complex of a mismatch and components of a mismatch repair system to other cellular proteins. Another aspect of the invention for removal of molecules containing a mismatch is through the incorporation of biotin into such a molecule and subsequent removal by binding to avidin.

Another aspect of the invention features use of a mismatch repair system which has a defined 5' to 3' exonuclease function, that is provided by the exonuclease VII or RecJ exonuclease. In other systems a 5' to 3' exonuclease function is provided by exonuclease VII which is present in many preparations of the DNA polymerase III holoenzyme.

The invention also includes kits having components necessary to carry out the methods of the invention.

The mismatch repair systems of the instant invention, e.g., *E. coli*, offer specific and efficient procedures for detection and localization of mismatches and manipulation of DNA containing mismatches that is a reflection of their biological function. All eight possible base pair mismatches are recognized and seven of the eight mismatches are processed and corrected by the system. Although C—C mismatches are not a substrate for repair, MutS does bind weakly to this mispair permitting its detection. In contrast to the electrophoretic migration procedure, the RNase method, or chemical modification procedures, the system does not depend on the destabilization of the DNA helix for detection of mismatches or binding to mismatches. The system features exquisite specificity, and is not subject to non-specific interactions with bases at the ends of linear DNA fragments or non-specific interactions at non-mismatch sites in long molecules.

The detection of fragments containing a mispair is limited only by the intrinsic specificity of the system, for example, detection of better than one G—T mispair per 300 kilobases. Mismatches have been routinely detected with a 6,400 base pair substrate and the system should be applicable to molecules as large as 40–50 kb. This allows for detection of possible single base differences between long DNA sequences, for example, between a complete gene from one individual and the entire genome of another. The invention also enables the localization of any possible single base

10

difference within the sequences of homologous regions of long DNA molecules such as those encoding one or more complete genes and comprising several kbp of DNA.

Several of the methods of the invention result in the covalent alteration of the phosphodiester backbone of DNA molecules. This covalent alteration facilitates analysis of the product DNA molecules especially by electrophoretic methods.

Other features and advantages of the invention will be apparent from the following description of the preferred embodiments thereof, and from the claims.

BRIEF DESCRIPTION OF THE FIGURES

FIG. 1. Heteroduplex substrate for in vitro mismatch correction. The substrate used in some examples is a 6440-bp, covalently closed, circular heteroduplex that is derived from bacteriophage f1 and contains a single base-base mismatch located within overlapping recognition sites for two restriction endonucleases at position 5632. In the example shown a G—T mismatch resides within overlapping sequences recognized by Hind III and Xho I endonucleases. Although the presence of the mispair renders this site resistant to cleavage by either endonuclease, repair occurring on the complementary (c) DNA strand yields an A—T base pair and generates a Hind III-sensitive site, while correction on the viral (v) strand results in a G—C pair and Xho I-sensitivity. The heteroduplexes also contain a single d(GATC) sequence 1024 base pairs from the mismatch (shorter path) at position 216. The state of strand methylation at this site can be controlled, thus permitting evaluation of the effect of DNA methylation on the strand specificity of correction.

FIG. 2. Requirement for DNA ligase in mismatch correction. Hemimethylated G—T heteroduplex DNA (FIG. 1, 0.6 µg, d (GATC) methylation on the complementary DNA strand) was subjected to mismatch repair under reconstituted conditions in a 60 µl reaction (Table 3, closed circular heteroduplex), or in 20 µl reactions (0.2 µg of DNA) lacking MutS protein or ligase, or lacking both activities. A portion of each reaction (0.1 µg of DNA) was treated with EDTA (10 mM final concentration) and subjected to agarose gel electrophoresis in the presence of ethidium bromide (1.5 µg/ml; top panel, lanes 1–4). Positions are indicated for the unreacted, supercoiled substrate (SC), open circles containing a strand break (OC) and covalently closed, relaxed circular molecules (RC). A second sample of each reaction containing 0.1 µg of DNA was hydrolyzed with Xho I and Cla I endonucleases (FIG. 1) to score G—T to G—C mismatch correction and subjected to electrophoresis in parallel with the samples described above (bottom panel, lanes 5–8). The remainder of the complete reaction (0.4 µg DNA, corresponding to the sample analyzed in lane 1) was made 10 mM in EDTA, and subjected to electrophoresis as described above. A gel slice containing closed circular, relaxed molecules was excised and the DNA eluted. This sample was cleaved with Xho I and Cla I and the products analyzed by electrophoresis (lane 9).

FIG. 3. Methyl-direction of mismatch correction in the purified system. Repair reactions with the G—T heteroduplex (FIG. 1) were performed as described in Table 3 (closed circular heteroduplex) except that reaction volumes were 20 µl (0.2 µg of DNA) and the incubation period was 60 minutes. The reactions were heated to 55° for 10 minutes and each was divided into two portions to test strand specificity of repair. G—T to A—T mismatch correction, in which repair occurred on the complementary (c) DNA

5,679,522

11

strand, was scored by cleavage with Hind III and Cla I endonucleases, while hydrolysis with Xho I and Cla I were used to detect G—T to G—C repair occurring on the viral (v) strand. Apart from the samples shown in the left two lanes, all heteroduplexes were identical except for the state of methylation of the single d(GATC) sequence at position 216 (FIG. 1). The state of modification of the two DNA strands at this site is indicated by + and - notation. The G—T heteroduplex used in the experiment shown in the left two lanes (designated 0/0) contains the sequence d(GATT) instead of d(GATC) at position 216, but is otherwise identical in sequence to the other substrates.

FIG. 4. Strand-specific repair of heteroduplexes containing a single strand scission in the absence of MutH protein. Hemimethylated G—T heteroduplex DNAs (FIG. 1, 5 µg) bearing d(GATC) modification on the vital or complementary strand were subjected to site-specific cleavage with near homogeneous MutH protein. Because the MutH-associated endonuclease is extremely weak in the absence of other mismatch repair proteins, cleavage at d(GATC) sites by the purified protein requires a MutH concentration 80 times that used in reconstitution reactions. After removal of MutH by phenol extraction, DNA was ethanol precipitated, collected by centrifugation, dried under vacuum, and resuspended in 10 mM Tris-HCl (pH 7.6), 1 mM EDTA. Mismatch correction of MutH-incised and covalently closed, control heteroduplexes was performed as described in the legend to Table 2 except that ligase and NAD⁺ were omitted. Outside and inside strands of the heteroduplexes depicted here correspond to complementary and viral strands respectively. Values in parentheses indicate repair occurring on the methylated, continuous DNA strand. The absence of MutH protein in preparations of incised heteroduplexes was confirmed in two ways. Preparations of incised molecules were subject to closure by DNA ligase (>80%) demonstrating that MutH protein does not remain tightly bound to incised d(GATC) sites. Further, control experiments in which each MutH-incised heteroduplex was mixed with a closed circular substrate showed that only the open circular form was repaired if MutH protein was omitted from the reaction whereas both substrates were corrected if MutH protein was present (data not shown).

FIG. 5. Requirements for MutH protein and a d(GATC) sequence for correction in the presence of DNA ligase. Hemimethylated G—T heteroduplexes incised on the unmethylated strand at the d(GATC) sequence were prepared as described above in FIG. 4. A G—T heteroduplex devoid of d(GATC) sites (FIG. 4) and containing a single-strand break within the complementary DNA strand at the Hinc II site (position 1) was constructed as described previously (Lahue et al. supra). Mismatch correction assays were performed as described in Table 3, with ligase (20 ng in the presence of 25 µM NAD⁺) and MutH protein (0.26 ng) present as indicated. Table entries correspond to correction occurring on the incised DNA strand, with parenthetical values indicating the extent of repair on the continuous strand. Although not shown, repair of the nicked molecule lacking a d(GATC) sequence (first entry of column 3) was reduced more than an order of magnitude upon omission of MutL, MutS, SSB or DNA polymerase III holoenzyme.

FIG. 6 is a diagrammatic representation of the model system used to evaluate MutS and MutL effects on RecA catalyzed strand transfer.

FIG. 7 depicts the effects of MutS and MutL on RecA-catalyzed strand transfer between homologous and quasi-homologous DNA sequences. Solid lines indicate fd—fd strand transfer, while dashed lines correspond to fd-M13

12

strand transfer. Strand transfer was evaluated in the presence of MutL (solid circles) or MutS (solid squares).

FIG. 8 depicts The MutL potentiation of MutS block to strand transfer in response to mismatched base pairs. Solid lines: fd-fd strand transfer; dashed lines fd-M13 strand transfer; RecA (solid circle); RecA and MutL (solid diamond); RecA and MutS (solid square); RecA, MutL, and MutS (solid triangle).

FIG. 9 depicts the MutS and MutL block of branch migration through regions that generate mismatched base pairs. Solid lines: M13-M13 strand transfer; dashed line fd-M13 strand transfer. RecA only (solid circle and square); RecA, MutS, and Mutn (open circle and square).

DESCRIPTION OF PREFERRED EMBODIMENTS

The invention consists of methods utilizing and kits consisting of components of mismatch repair system to detect, and localize DNA base pair mismatches and manipulate molecules containing such mismatches. The invention also features modified mispair recognition proteins and their utilization in the above-mentioned methods and kits. The invention also includes methods and kits comprising components of a mismatch repair system along with proteins with recombinase activity. The invention also consists of methods to improve the GMS technique to detect regions of homology-by-descent.

Methods for detecting the presence and localization of mismatched base pairs by complex formation with a mismatch recognition protein

One embodiment of the invention features a diagnostic method for detecting a base pair mismatch in a DNA duplex. The method comprises the steps of contacting at least one strand of a first DNA molecule with the complementary strand of a second DNA molecule under conditions such that base pairing occurs, contacting a DNA duplex potentially containing a base pair mismatch with a mispair recognition protein under conditions suitable for the protein to form a specific complex only with the DNA duplex having a base pair mismatch, and not with a DNA duplex lacking a base pair mismatch, and detecting the complex as a measure of the presence of a base pair mismatch in the DNA duplex.

By "mismatch" is meant an incorrect pairing between the bases of two nucleotides located on complementary strands of DNA, i.e., bases pairs that are not A:T or G:C.

In the practice of this method, the two DNA's or two DNA samples to be compared may comprise natural or synthetic sequences encoding up to the entire genome of an organism, including man, which can be prepared by well known procedures. Detection of base sequence differences according to this method of this invention does not require cleavage (by a restriction nuclease, for example) of either of the two DNAs; although it is well known in the art that rate of base pair formation between complementary single-stranded DNA fragments is inversely related to their size. This detection method requires that base sequence differences, to be detected as base pair mismatches lie within a region of homology constituting at least about 14 consecutive base pairs of homology between the two DNA molecules, which is about the minimum number of base pairs generally required to form a stable DNA duplex. Either one or both of the strands of the first DNA may be selected for examination, while at least one strand of the second DNA complementary to a selected first DNA strand must be used. The DNA strands, particularly those of the second DNA, advantageously may be radioactively labeled to facilitate direct detection, according to procedures well known in the art.

5,679,522

13

By "mismatch recognition protein" is meant a protein of a mismatch repair system that specifically recognizes and binds to a base pair mismatch, e.g. *E. coli* MutS.

Methods and conditions for contacting the DNA strands of the two DNAs under conditions such that base pairing occurs are also widely known in the art.

In preferred embodiments of this aspect of this invention, the mismatch recognition protein is the product of the mutS gene of *E. coli*, or species variations thereof, or portions thereof encoding the recognition domain. The protein recognizes all eight possible base pair mismatches, detection of the DNA:protein complex comprises contacting the complexes with a selectively absorbent agent under conditions such that the protein:DNA complexes are retained on the agent while DNA not complexed with protein is not retained and measuring the amount of DNA in the retained complexes, the absorbent agent is a membranous nitrocellulose filter, detection of the DNA:protein complex further includes the step wherein an antibody specific for the base mismatch recognition protein is employed, the base mismatch recognition protein is the product of the mutS gene of *S. typhimurium* the hexA gene of *S. pneumoniae* or the MSH1 and MSH2 genes of yeast, and wherein the step for detecting the DNA:protein complex further includes a step wherein the electrophoretic mobility of the DNA:protein complex is compared to uncomplexed DNA.

The ability of the MutS protein to recognize examples of all eight single base pair mismatches within double-stranded DNA, even including C—C mismatches which do not appear to be corrected in vivo, is demonstrated by the fact that MutS protein protects DNA regions containing each mismatch from hydrolysis by DNase I (i.e., by "Dnase I footprint" analyses), as recently reported (Su, S. -S., et al., 1988, *J. Biol. Chem.*, 263, 6829–6835). The affinity of MutS protein for the different mismatches that have been tested varies considerably. Local sequence environment may also affect the affinity of the MutS protein for any given base mismatch; in other words, for example, the affinity for two specific cases of A—C mismatches, which are surrounded by different sequences, may not be the same. Nevertheless, no examples of base mismatches have been found that are not recognized by isolated MutS protein. Accordingly, this method of the invention detects all mismatched base pairs.

It should be particularly noted that the DNA duplexes which MutS recognizes are not required to contain GATC sequences and, hence, they do not require hemimethylation of A's in GATC sequences, the specific signal for the full process of methyl-directed mismatch correction in vivo; therefore, use of MutS in this method allows recognition of a DNA base mismatch in DNAs lacking such methylation, for instance, DNAs isolated from human tissues.

By "species variation" is meant a protein which appears to be functionally and in part, at least, structurally homologous to the *E. coli* MutS protein. One example of such a protein has also been discovered in a methyl-directed mismatch correction system in *Salmonella typhimurium* bacteria (Pang et al., 1985, *J. Bacteriol.*, 163, 1007–1015). The gene for this protein has been shown to complement *E. coli* strains with mutations inactivating the mutS gene and the amino acid sequence of its product shows homology with that of the *E. coli* MutS protein. Accordingly, this *S. typhimurium* protein is also suitable for the practice of this aspect of the present invention. Other organisms, including man, are known to possess various systems for recognition and repair of DNA mismatches, which, as one skilled in the art would appreciate, comprise mismatch recognition proteins functionally homologous to the MutS protein. Nuclear extracts

14

derived from HeLa and *Drosophila melanogaster* K_o cell lines has been shown to support efficient strand-specific mismatch correction in vitro (Holmes et al., 1990, *Proc. Natl. Acad. Sci. USA* 87, 5837–5841, which is incorporated herein by reference), and this reaction has been shown to occur by a mechanism similar to that of the bacterial reaction (Fany and Modrich 268 *J. Biol. Chem.* 11838, 1993). Furthermore, genes encoding proteins that are homologous to bacterial MutS at the amino acid sequence level have been demonstrated in human (Fujii and Shimada 264 *J. Biol. Chem.* 10057, 1989) and yeast (Reenan and Kolodner 132 *Genetics* 963, 1992) and *S. pneumoniae* (Priebe et al., 170 *J. Bacteriol.* 190, 1988). Accordingly, it is believed that such DNA base mismatch recognition proteins may also be suitable for use in the present invention.

By "protein encoding the recognition domain" is meant a region of the mismatch recognition protein which is involved in mismatch recognition and binding. Such a domain comprises less than the complete mismatch recognition protein.

By a "selectively adsorbent agent" is meant any solid substrate to which protein:DNA complexes are retained on the agent while DNA not complexed with protein is not retained, such agents are known to those skilled in the art. Absent radioactive labeling of at least one strand used to form the DNA duplexes, the DNA in complexes on the filter may be detected by any of the usual means in the art for detection of DNA on a solid substrate, including annealing with complementary strands of radioactive DNA.

The nitrocellulose filter method for detecting complexes of MutS protein with base mismatches in DNA has been reported in detail (Jiricny, J. et al., 1988, *Nuc. Acids Res.* 16, 7843–7853, which is hereby incorporated herein by reference). Besides simplicity, a major advantage of this method for detecting the DNA:protein complex over other suitable methods is the practical lack of a limitation on the size of DNA molecules that can be detected in DNA:protein duplexes. Therefore, this embodiment of this method is in principle useful for detecting single base sequence differences between DNA fragments as large as can be practically handled without shearing.

By "electrophoretic mobility" is meant a method of separating the DNA:protein complexes from DNA that does not form such complexes on the basis of migration in a gel medium under the influence of an electric field. DNA:protein complexes are less mobile than naked DNA. Such methods based on electrophoretic mobility are known to those skilled in the art. The DNA in the DNA:protein complexes may be detected by any of the usual standard means for detection of DNA in gel electrophoresis, including staining with dyes or annealing with complementary strands of radioactive DNA. Detecting complexes comprising the MutS base mismatch recognition protein and mismatches in DNA duplexes is also described in the foregoing reference (Jiricny, J. et al., 1988, *Nuc. Acids Res.*, 16, 7843–7853). Under the usual conditions employed in the art for detecting specific DNA:protein complexes by gel electrophoresis, complex formation of a protein with a double-stranded DNA fragment of up to several hundred base pairs is known to produce distinguishable mobility differences.

Antibodies specific for a DNA mismatch recognition protein can be prepared by standard immunological techniques known to those skilled in the art.

Other suitable analytical methods for detecting the DNA protein complex include immunodetection methods using an antibody specific for the base mismatch recognition protein. For example, antibodies specific for the *E. coli* MutS protein have been prepared. Accordingly, one immunodetection

5,679,522

15

method for complexes of MutS protein with DNA comprises the steps of separating the DNA:protein complexes from DNA that does not form such complexes by immunoprecipitation with an antibody specific for MutS protein, and detecting the DNA in the precipitate. According to the practice of this aspect of the invention, quantitative immunoassay methods known in the art may be employed to determine the number of single base mispairs in homologous regions of two DNA molecules, based upon calibration curves that can be established using complexes of a given mispair recognition protein with DNA duplexes having known numbers of mispairs.

Another aspect of the invention features a method for detecting and localizing a base pair mismatch in a DNA duplex. The method includes contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting double-stranded DNA duplexes with a mispair recognition protein under conditions such that the protein forms specific complexes with mispairs, subjecting the duplex molecules to hydrolysis with an exonuclease under conditions such that the complex blocks hydrolysis, and determining the location of the block to hydrolysis by a suitable analytic method.

"Hydrolysis with an exonuclease" is a procedure known to those skilled in the art and utilizes enzymes possessing double-strand specific exonuclease activity, e.g., *E. coli* exonuclease III, RecBCD exonuclease, lambda exonuclease, and T7 gene 6 exonuclease.

By "block to hydrolysis" is meant interference of hydrolysis by the exonuclease. Such protection can result from the mispair recognition protein protecting the DNA to which it is bound.

By "suitable analytical method" is meant any method that allows detection of the block to exonuclease digestion, such analysis of molecules by gel electrophoresis. Such methods are known to those skilled in the art.

Methods for detecting and localizing base pair mismatches by mismatch repair system strand modification reactions

In addition to methods that detect base sequence differences, this invention provides methods for both detecting and localizing a base pair mismatch in a DNA duplex. One method includes contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting double-stranded DNA duplexes with a mismatch recognition protein under conditions such that the protein forms specific complexes with mispairs and thereby directs modification of at least one strand of the DNA in the resulting DNA:protein complexes in the vicinity of the DNA:protein complex, and determination of the location of the resulting DNA modification by a suitable analytic method.

By "modification" is meant any alteration for which there is a means of detection, for instance a chemical modification including breaking of a chemical bond resulting in, as examples, cleavage between nucleotides of at least one DNA strand or removal of a base from the sugar residue of a nucleotide. Specific means for modifying DNAs in the vicinity of the DNA:protein complex are provided below for several embodiments of this aspect of the invention, together with interpretations of the phrase "in the vicinity of", as appropriate to the practical limitations of the modification approach in each instance.

Suitable analytical methods for determining the location of the modification are known to those skilled in the art. Such a determination involves comparison of the modified DNA molecule with the homologous unmodified DNA molecule.

16

In preferred embodiments of this aspect of the invention, the mispair recognition protein is the product of the mutS gene of *E. coli* or another functionally homologous protein; the step in which the DNA is modified in the vicinity of the DNA:protein complex further comprises contacting the DNA:MutS protein complex with a defined set or subset of *E. coli* DNA mismatch repair proteins (comprising *E. coli* Muth, Mutn, DNA helicase II, single-stranded DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, and exonuclease VII (or RecJ exonuclease), or species variations of these activities), ATP and one or more dideoxynucleoside-5'-triphosphates or in the absence of exogenous deoxyribonucleoside-5'-triphosphate under conditions that produce a discontinuity in one or both strands of the DNA duplex in the vicinity of the mismatch.

DNA used in such an analysis are to be unmethylated or hemimethylated at on the 6-position of the adenine base in GATC sequences. With the exception of DNAs from some bacterial species, the chromosomes of most organisms naturally lack this modification. In those cases where hemimethylation of otherwise GATC unmodified molecules is desired, this can be accomplished by use of *E. coli* Dam methylase as is well known in the art. Symmetrically methylated DNA prepared by use of this enzyme is denatured and subsequently reannealed with single-stranded sequences representing an homologous (or largely so) DNA. If necessary, hemimodified molecules produced by this renaturation procedure can be separated from unmethylated is symmetrically methylated duplexes which can also result from the annealing procedure. As is well known in the art, this can be accomplished by subjecting annealed products to cleavage by DpnI and MboI endonucleases. The former activity cleaves symmetrically methylated duplex DNA at GATC sites while unmodified duplex DNA is subject to double strand cleavage only at unmodified GATC sites by the latter activity. Since hemimodified DNA is resistant to double strand cleavage by both DpnI and MboI, desired hemimethylated products can be separated on the basis of size from the smaller fragments produced by DpnI and MboI cleavage, for example by electrophoretic methods.

By "discontinuity in one or both strands of the DNA duplex" is meant a region which consists of a break in the phosphodiester backbone in one or both strands, or a single-stranded gap in a duplex molecule.

One aspect of this preferred embodiment involves contacting the DNA:MutS protein complex with *E. coli* MutL and Muth proteins (or species variations thereof) in the presence of ATP and an appropriate divalent cation cofactor (eg, Mg^{2+}) so that mismatch-containing molecules will be subject to incision at one or more GATC sites in the vicinity of the mispair. Such incision events can be monitored by a suitable analytic method for size detection such as electrophoresis under denaturing condition.

A second aspect of this preferred embodiment involves contacting the DNA:MutS complex with a defined *E. coli* mismatch correction system consisting of *E. coli* Muth, MutL, DNA helicase II, single-stranded DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, and exonuclease VII (or RecJ exonuclease), or species variants of these activities, ATP in the absence of exogenous deoxyribonucleoside-5'-triphosphates or in the presence of one or more dideoxynucleoside-5'-triphosphates such that single-stranded gaps are produced in the vicinity of the complexed protein; the method for determining the location of the single-stranded gaps with the DNA duplex further includes analysis of electrophoretic mobility of treated samples under denaturing conditions of the steps of cleaving

5,679,522

17

the DNA with a single-stranded specific endonuclease, and comparing the electrophoretic mobilities of the cleaved fragments with unmodified DNA fragments under non-denaturing conditions; the step for modifying the DNA duplex in the vicinity of the complexed protein comprises contacting the complexes with proteins of a mismatch repair system, ATP and a divalent cation under conditions such that an endonucleolytic incision is introduced at one or more GATC sequences in the duplex molecule.

An example of a complete defined mismatch correction system comprises the following purified components: *E. coli* MutH, MutL, and MutS proteins, DNA helicase II, single-stranded DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, DNA ligase, ATP, and the four deoxynucleoside-5'-triphosphates. This set of proteins can process seven of the eight base-base mismatches in a strand-specific reaction that is directed by the state of methylation of a single GATC sequence located 1 kilobase from the mispair. This defined system is described further in Example 1, below. The 5' to 3' exonuclease function can either be supplied by either DNA polymerase III holoenzyme preparations that contain this activity or as a separate defined component consisting of exonuclease VII or RecJ exonuclease. It should be noted that the lack of ability to repair C—C base mispairs in this embodiment of this aspect of the present invention is not a major limitation of the method for detecting all possible base sequence differences between any two naturally occurring DNA sequences because mutations that would give rise to a C—C mispair upon hybridization would also give rise to a G—G mismatch when the complementary strands are hybridized.

For the purpose of generating single-stranded gaps in the vicinity of the DNA:MutS protein complexes, DNA duplexes containing mispaired base pairs are contacted with the defined mismatch correction system under the standard conditions described in Example 1, Table 3 (Complete reaction), except for the following differences: (i) exogenous dNTPs are omitted; or (ii) 2', 3'-dideoxynucleoside-5'-triphosphates (ddNTPs) at suitable concentrations (10 to 100 μ M) are substituted for dNTPs; or (iii) reactions containing dNTPs are supplemented with ddNTPs at a suitable concentration to yield a chain termination frequency sufficient to inhibit repair of single-strand gaps. In cases (i)–(iii) DNA ligase may be omitted from the reaction. In cases (ii) and (iii) all four ddNTPs may be present; however, it is expected that the presence of one, two, or three ddNTPs will prove sufficient to stabilize single strand gaps via chain termination events. While it is expected that most applications of these gap forming protocols will utilize MutH, it is pertinent to note that the requirement of methyl-direct strand incision by MutH may be obviated by provision of a single-strand nick by some other means within the vicinity of the mispair, as described in Example 1, FIG. 5. A suitable means for inducing such nicks in DNA is limited contact with a nuclease, Dnase I, for example; under conditions that are well known in the art, this approach creates nicks randomly throughout double-stranded DNA molecules at suitable intervals for allowing the mispair correction system to create single-stranded gaps in the vicinity of a mispair anywhere in the DNA.

It should be noted that in this embodiment of this method for localizing mismatch base pairs, "in the vicinity of" a base mispair is defined practically by the size of the single-strand gaps typically observed under above conditions, namely up to about one kbp from the mismatched base pair.

By "determining the location of the single-stranded gaps within the DNA duplex" entails the steps of: (i) Cleaving the

18

DNA with at least one restriction endonuclease (either prior or subsequent to contact of the preparation with mismatch repair activities) followed by comparison of electrophoretic mobilities under denaturing conditions of the resulting modified DNA fragments with DNA restriction fragments not contacted with the defined mismatch correction system; or (ii) Cleaving the DNA with at least one restriction endonuclease and with a single-strand specific endonuclease, followed by comparison of the electrophoretic mobilities under native conditions of the resulting modified DNA fragments with DNA restriction fragments not contacted with the defined mismatch correction system. Suitable single-strand specific endonucleases include the S1 single-stranded specific nuclease, for example, or other functionally similar nucleases well known in the art. In the cases of either (i) or (ii), additional restriction mapping may be performed as needed to further localize any fragment modifications observed in initial application of the method, until, if desired, a restriction fragment of convenient size for direct sequence determination is obtained for direct comparisons of sequences of the two DNA molecules in the vicinity of the base sequence difference.

By "proteins of a mismatch repair system" are meant a protein that contains a GATC endonuclease, a mispair recognition protein, and proteins that participate in the activation of the GATC endonuclease.

By "divalent cation": is meant a cofactor for the GATC endonucleases, e.g., $MgCl_2$.

By "endonucleolytic incision": is meant cleavage of a DNA fragment containing a mismatched base pair at unmethylated or hemimethylated GATC sequences in the vicinity of a mismatch.

"Size fractionation by electrophoretic mobility under denaturing conditions" is a procedure well known by those skilled in the art. Gel Electrophoresis can either be conventional or pulse-field.

Modification of mispair recognition proteins and uses

The present invention also includes forms of mispair recognition proteins which have been altered to provide means for modifying at least one strand of the DNA duplex in the vicinity of the bound mispair recognition protein.

In preferred embodiments of this aspect of the invention, the altered mispair recognition protein is the modified product of the mutS gene of *E. coli* or is another functionally homologous modified protein to which is attached an hydroxyl radical cleaving function; the altered mispair recognition protein may comprise only a segment of the native molecule containing the mispair recognition domain; the hydroxyl radical cleaving function is selected from the group consisting of the altered mispair recognition protein wherein the hydroxyl radical cleaving function is selected from the group consisting of the 1,10-phenanthroline-copper complex, the EDTA iron complex, and the copper binding domain of serum albumin; the altered mispair recognition protein is the product of the mutS gene of *E. coli* or of another functionally homologous protein to which is attached attachment a DNA endonuclease activity capable of cleaving double-stranded DNA; the endonuclease activity is provided by the DNA cleavage domain of FokI endonuclease.

By "altered mispair recognition protein" is meant a mispair recognition protein that not only recognizes and binds to a base pair mismatch, but possess the ability to modify a strand of a DNA molecule containing such a mismatch.

Several methods for attaching an hydroxyl radical cleaving function to a DNA binding protein are known in the art. For example, lysyl residues may be modified by chemically

5,679,522

19

attaching the 1,10-phenanthroline-copper complex to lysine residues, resulting in conversion of a DNA binding protein into a highly efficient site-specific nuclease that cleaved both DNA strands (in the presence of hydrogen peroxide as a coreactant) within the 20 base pair binding site of the protein, as determined by DNase I footprinting (C. -H. Chen and D. S. Sigman, 1987, *Science*, 237, 1197). Chemical attachment of an EDTA-iron complex to the amino terminus of another DNA binding protein similarly produced a sequence specific DNA cleaving protein that cut both strands of the target DNA within a few bases of recognition site of similar size (J. P. Sluka, et al., 1987, *Science*, 235, 777).

An alternate means for attaching the hydroxyl radical cleaving function to this same protein involved extension of the amino terminus with the three amino acids, Gly—Gly—His, which is consensus sequence for the copper-binding domain of serum albumin (D. P. Hack et al., 1988, *J. Am. Chem. Soc.*, 110, 7572–7574). This approach allows for preparation of such an artificial DNA cleaving protein directly by recombinant methods, or by direct synthesis using standard solid phase methods, when the peptide is sufficiently short as it was in this case (55 residues including the 3 added amino acids), thereby avoiding the need for an additional chemical modification step of the reagent which is both time consuming and difficult in large scale production. In contrast to the EDTA-iron complex, the particular peptide sequence constructed in this instance cleaved only one example out of four recognition sites in different sequence environments.

Nevertheless, one skilled in the art of protein engineering would appreciate that this general approach for converting a DNA binding protein into a DNA cleaving protein by attachment of an hydrogen radical cleavage function is widely applicable. Hence, DNA base mismatch recognition proteins which normally only bind to DNA are modified to cleave DNA by attachment of an hydroxyl radical cleavage function, according to the practice of this aspect of this invention, without undue experimentation, by adjustment of appropriate variables taught in the art, particularly the chemical nature and length of the "spacer" between the protein and the metal binding site.

Additional altered forms of mismatch recognition proteins that modify at least one strand of the DNA in a DNA:protein complex in the vicinity of the bound protein according to the present invention include proteins comprising the portions or "domains" of the unmodified base mismatch recognition enzymes that are essential for binding to a DNA mismatch. These essential DNA binding domains further comprise peptide sequences that are most highly conserved during evolution; such conserved domains are evident, for example, in comparisons of the sequences of the *E. coli* MutS protein with functionally homologous proteins in *S. typhimurium* and other structurally similar proteins. Accordingly, peptide sequences of a DNA base mismatch recognition protein that are protected from proteases by formation of specific complexes with mismatches in DNA and, in addition or in the alternative, are evolutionarily conserved, form the basis for a particularly preferred embodiment of this aspect of the present invention, since such peptides constitute less than half the mass of the intact protein and, therefore, are advantageous for production and, if necessary, for chemical modification to attach a cleavage function for conversion of the DNA binding protein into a DNA cleavage protein specific for sites of DNA base mismatches.

The DNA cleavage domain of FokI endonuclease has been defined (Li et al, 1992, *Proc. Natl. Acad. Sci. U.S.A.*, 89:4275).

20

Another embodiment of this aspect of the invention consists of a method for detecting and localizing a base pair mismatch within a DNA duplex, including the steps of contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs; contacting resultant duplex DNA molecules with an altered mismatch recognition protein, under conditions such that the protein forms specific complexes with a mismatch and thereby directs modification of at least one strand of the DNA in the resulting DNA protein complexes in the vicinity of the DNA protein complex, and determining the location of the modification of the DNA by a suitable analytic method.

In the detection and localization of a base pair mismatch method according to this embodiment which employs an altered mismatch recognition protein, and the modification comprises double-stranded cleavage of the DNA within the vicinity of any base mismatch wherein the "vicinity" substantially corresponds to the sequence of DNA protected by the binding of the protein to a base mismatch, generally within about 20 base pairs. A single-strand specific nuclease, S1, for instance, may be used to augment cleavage by the modified base mismatch recognition protein in the event that a single-strand bias is suspected in the cleavage of any DNAs with which the protein forms a specific complex. Alternatively, DNA's subject to cleavage by the modified mismatch recognition protein may be analyzed by electrophoresis under denaturing conditions. Location of the modification is by suitable analytical methods known to those skilled in the art. Methods utilizing mismatch repair systems to detect A—G base pair mismatches

In a preferred embodiment, a method for detecting and localizing A—G mismatches in a DNA duplex, includes the steps of contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs; contacting resultant duplex DNA molecules with a mismatch recognition protein that recognizes A—G mismatches and an apurinic endonuclease or lyase under conditions such that in the presence of a mismatch an endonucleolytic incision is introduced in the duplex molecule, and determining the location of the incision by a suitable analytic method.

In preferred embodiments the A—G mismatch recognition protein is the product of the mutY gene of *E. coli*; and the analytical method includes gel electrophoresis.

The present invention also comprises DNA mismatch recognition protein that recognizes primarily A—G mismatches without any apparent requirement for hemimethylation. One example of this protein is the product of the mutY gene of *E. coli*, is a glycosylase which specifically removes the adenine from an A—G mismatch in a DNA duplex. The MutY protein has been purified to near homogeneity by virtue of its ability to restore A—G to C—G mismatch correction to cell-free extracts (K. G. Au et al., *Proc. Nat. Acad. Sci. U.S.A.*, 85, 9163, 1988) of a mutS mutY double mutant strain of *E. coli*, as described in Example 2, below. Its electrophoretic migration in the presence of dodecyl sulfate is consistent with a molecular weight of 36 kDa, and it apparently exists as a monomer in solution. MutY, an apurinic (AP) endonuclease, DNA polymerase I, and DNA ligase are sufficient to reconstitute MutY-dependent, A—G to C—G repair in vitro. A DNA strand that has been depurinated thusly by the MutY protein is susceptible to cleavage by any of several types of AP endonuclease or lyase (e.g., human AP endonuclease II) or by piperidine, under conditions that are well known in the art. The cleavage products are then analyzed by gel electrophoresis under denaturing

5,679,522

21

conditions. Accordingly, this MutY protein is useful in a method for the specific detection and localization of A—G mispairs, according to the practice of the present invention, and hence identification of A•T to C•G or G•C to T•A mutations.

Sources of DNA fragments to be analyzed

In another embodiment of the invention, DNA molecules are obtained from the following sources: different individuals of the same species, individuals of different species, individuals of different kingdoms, different tissue types, the same tissue type in different states of growth, different cell types, cells of the same type in different states of growth, and cells of the same origin in different stages of development, and cells of the same type that may have undergone differential somatic mutagenesis, e.g., one class of which may harbor per cancerous mutation(s).

In a preferred embodiment, the DNA molecules comprise a probe sequence that has been at least partially characterized.

By "probe sequence that has been at least partially characterized" is meant a DNA molecule from any source that has been characterized by restriction mapping or sequence analysis, such techniques are known to those skilled in the art.

Kits comprising a mispair recognition protein

Another aspect of the invention features assay kits designed to provide components to practice the methods of the invention.

In one aspect the invention features an assay kit for detecting a base pair mismatch in a DNA duplex. The kit comprises one or more of the following components: an aliquot of a mispair recognition protein, an aliquot of control oligonucleotides, and an exonuclease.

In a preferred embodiment the mispair recognition protein is the product of the mutS gene of *E. coli*.

By "control oligonucleotides" is meant oligonucleotides for assaying the binding of the mismatch repair protein to a base pair mismatch. One set of oligonucleotides are perfectly homologous (negative control) and thus are not bound by the mispair recognition protein. Another set of oligonucleotides containing a base pair mismatch (positive control) and thus are bound by the mispair recognition protein.

By "exonuclease" is meant enzymes possessing double-strand specific exonuclease activity, e.g. *E. coli* exonuclease III, RecBCD exonuclease, lambda exonuclease, and T7 gene 6 exonuclease.

Another aspect of the invention features an assay kit for detecting and localizing a base pair mismatch in a DNA duplex. The kit comprises one or more of the following components: an aliquot of all or part of a mismatch repair system, an aliquot of dideoxynucleoside triphosphates; and a single-strand specific endonuclease.

By "all or part of a mismatch repair system" is meant either the complete system which is capable of repairing a base pair mismatch, for example, the three *E. coli* proteins MutH, MutL, and MutS, DNA helicase II, single-strand binding protein, DNA polymerase III, exonuclease I, exonuclease VII or RecJ exonuclease, DNA ligase and ATP, or only the three proteins MutH, MutL, and MutS, along with ATP such that an endonucleolytic incision is made at a GATC site, with no subsequent repair reaction taking place.

In preferred embodiments the mismatch repair system includes: the products of the *E. coli* mutH, mutL, and mutS genes, or species variations thereof, DNA helicase II, single-strand DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, exonuclease VII or RecJ

22

exonuclease, DNA ligase, and ATP, the mismatch repair system includes only the products of the *E. coli* mutH, mutL, and mutS genes, or species variations thereof, and ATP.

Another embodiment of the invention features an assay kit for detecting and localizing a base pair mismatch in a DNA duplex comprising an aliquot of a modified mispair recognition protein.

In a preferred embodiment the mispair recognition protein is the product of the mutS gene of *E. coli*.

A further embodiment of this aspect of the invention features an assay kit for detecting and localizing an A—G mispair within a DNA duplex. The kit comprises one or more of the following components: an aliquot of an A—G mispair recognition protein; and an aliquot of an apurinic endonuclease or lyase.

In a preferred embodiment the A—G mispair recognition protein is the product of the MutY gene of *E. coli*.

Methods utilizing mismatch repair systems and recombinase proteins

In a further aspect, the invention features a method for eliminating DNA molecules containing one or more mismatches from a population of heterohybrid duplex DNA molecules formed by base pairing of single-stranded DNA molecules obtained from a first source and a second source.

The method includes digesting genomic DNA from the first and the second source with a restriction endonuclease, methylating the DNA of one of the sources, denaturing the DNA from one or both sources, mixing the DNA molecules from the first and the second source in the presence of a recombinase protein, proteins of a mismatch repair system that modulate the recombinase protein, single-strand binding protein, and ATP under conditions such that DNA duplexes form in homologous regions of the DNA molecules from the first and the second source and the presence of a base pair mismatch results in regions that remain single-stranded, and removing molecules that contain single-stranded regions from the population.

By "heterohybrid" is meant a duplex DNA molecule that consists of base-paired strands originating from two different sources, such that one strand of the duplex is from one source (first source) and the other strand is from another source (second source).

The "source" of DNA molecules designates the origin of the genomic DNA used in the method. The first and second sources are different, i.e., not from the same cell of the same individual.

By "restriction endonuclease" is meant an enzyme which recognizes specific sequences in double-stranded DNA and introduces breaks the phosphodiester backbone of both strands. For use in the current invention restriction endonucleases that digest genomic DNA or cDNA into fragments of approximately 4 to 20 kilobases are preferred.

By "methylating" is meant the process by which a methyl groups is attached to the adenine residue of the sequence "GATC". This reaction is carried by enzymes well known in the art, such as the DAM system of *E. coli*.

By "denaturing" is meant the process by which strands of duplex DNA molecules are no longer based paired by hydrogen bonding and are separated into single-stranded molecules. Methods of denaturation are well known to those skilled in the art and include thermal denaturation and alkaline denaturation.

By "recombinase protein" is meant a protein that catalyzes the formation of DNA duplex molecules. Such a molecule is capable of catalyzing the formation of duplex DNA molecules from complementary single-stranded molecules by renaturation or by catalyzing a strand transfer

5,679,522

23

reaction between a single-stranded molecule and a double-stranded molecule. Examples of such a protein are the RecA proteins of *E. coli* and *S. typhimurium*.

By "proteins of a mismatch repair system that modulate the recombinase protein" are meant components of a system which recognizes and corrects base pairing errors in duplex DNA molecules and also influence the activity of a recombinase protein. For example, a mispair recognition protein, e.g., MutS, and a protein that interacts with the mismatch repair protein, e.g., MutL, together inhibit duplex formation catalyzed by the recombinase protein in the presence of a base pair mismatch. Such modulation of the recombinase protein results in single-stranded regions downstream of the base pair mismatch.

In preferred embodiments, the recombinase protein is the *E. coli* RecA protein, the mismatch repair system is from *E. coli* and the components are the MutS and MutL proteins, the sources of DNA are different individuals of the same species, individuals of different species, individuals of different kingdoms, different tissue types, the same tissue type in different states of growth, different cell types, cells of the same type in different states of growth, cells of the same origin in different stages of development, and cells of the same origin that may have undergone differential somatic mutagenesis, the method of removing molecules containing single-stranded regions is by chromatography on benzoylated naphthoylated DEAE, the method of removing molecules containing single-stranded regions is by treatment with a single-strand specific nuclease.

The MutS, MutL protein along with single-strand binding protein and ATP are involved in modulation of the *E. coli* RecA protein in catalyzing heteroduplex formation.

The method for removing molecules containing single-strands from double-stranded molecules by the use of chromatography with benzoylated naphthoylated DEAE is well known to those skilled in the art.

By "single strand specific nuclease" is meant an enzyme that specifically degrades single-stranded regions of DNA molecules and do not degrade double stranded regions. Examples of such nucleases are: S1, mung bean, T7 gene 3 endonuclease and P1 nuclease.

In another aspect, the invention features a method for eliminating DNA molecules containing one or more mismatches from a population of heterohybrid duplex DNA molecules formed by a strand transfer reaction between duplex DNA molecules obtained from a first source and denatured DNA molecules from a second source. The method includes digesting genomic DNA from the first and the second source with a restriction endonuclease, methylating the DNA of one of the sources, denaturing the DNA from the second source, mixing the DNA molecules from the first and the second source in the presence of a protein which catalyzes strand transfer reactions, proteins of a mismatch repair system that modulate the protein with strand transfer activity, single strand binding protein, and ATP under conditions such that DNA heteroduplexes form in homologous regions of the DNA molecules from the first and the second source by strand transfer reaction and the presence of a base pair mismatch results in regions that remain single-stranded, and removing molecules that contain the single-stranded regions from the population.

By "strand transfer reaction is meant" a three strand reaction between duplex DNA from one source and single-stranded DNA from another source in which one strand of the duplex is displaced by a single-stranded molecule.

By "a protein which catalyzes strand transfer reaction" is meant proteins such as: RecA, homologs of RecA, and

24

proteins with branch migration enhancing activities such as RuvA, RuvB, RecG.

In preferred embodiments, the strand transferase protein is the *E. coli* RecA protein, the mismatch repair system is from *E. coli* and the components are the MutS and MutL proteins, the sources are different individuals of the same species, individuals of different species, individuals of different kingdoms, different tissue types, the same tissue type in different states of growth, different cell types, cells of the same type in different states of growth, and cells of the same origin in different stages of development, cells of the same origin that may have undergone differential somatic mutagenesis (e.g. normal as opposed to pre-tumor cells), a probe sequence that has been at least partially characterized, the method of removing molecules containing single-stranded regions is by chromatography on benzoylated naphthoylated DEAE, the method of removing molecules containing single-stranded regions is by treatment with a single strand specific nuclease.

Methods of improving the Genomic Mismatch Scanning technique

In another aspect the invention features the utilization of a recombinase or strand transferase and proteins of a mismatch repair system that modulate the recombinase or strand transferase, in the hybridization step of the genomic mismatch scanning technique. Formation of duplex molecules catalyzed by a recombinase or strand transferase protein which is modulated by components of a mismatch repair system, provide an additional selection step in the GMS method.

By "genomic mismatch scanning" is meant a technique to identify regions of genetic identity between two related individuals. Such a technique has been described by Nelson et al, 4 *Nature Genetics* 11, 1993.

In a further embodiment the invention features a method of genomic mismatch scanning such that heterohybrid DNA molecules containing a base pair mismatch are removed, without the use of exonuclease III. The method comprises the steps of contacting a population of heterohybrid DNA molecules potentially containing base pair mismatches with all the components of a DNA mismatch repair system in the absence of dNTP's or in the presence of one or more dideoxy nucleoside triphosphates under conditions such that single-stranded gaps are generated in DNA fragments that contained a base pair mismatch and removing the molecules containing single-stranded gaps.

In preferred embodiments the DNA mismatch repair system is the *E. coli* methyl-directed mismatch repair system; removal of molecules containing single-stranded regions is by chromatography on benzoylated naphthoylated DEAE; removal of molecules containing single-stranded regions is by treatment with a single-strand specific nuclease.

In a further embodiment, the invention features another variation of the method of genomic mismatch scanning such that heterohybrid DNA molecules containing base pair mismatches are removed, without the use of exonuclease III. The method comprises the steps of contacting a population of heterohybrid DNA molecules potentially containing base pair mismatches with all the components of a DNA mismatch repair system and biotinylated nucleoside triphosphates under conditions such that biotinylated nucleotides are incorporated into DNA fragments that contained a base pair mismatch and, removing the molecules containing biotinylated molecules by binding to avidin.

Substitution with biotinylated nucleotides and binding of molecules that have incorporated these nucleotides are pro-

5,679,522

25

cedures well known to those skilled in the art. This procedure allows fractionation of a population of hybrid DNA molecules into two fractions: (i) A mismatch free fraction which fails to adhere to avidin; and (ii) A population that originally contained mispairs and which binds to avidin. The former can be utilized in the GMS procedure. The latter, avidin-bound class can be employed for other purposes. For example, when prepared using heterohybrid DNA produced by annealing DNA from two related haploid organisms the biotinylated sequences correspond to those DNA regions that vary genetically between the two organisms. Such sequences can thus be applied to determination of the molecular basis of genetic variation of organisms in question, e.g. pathogenic versus nonpathogenic microbial subspecies.

In a preferred embodiment the mismatch repair system is the methyl-directed mismatch repair system of *E. coli*.

In a further embodiment, the invention features a method of genomic mismatch scanning such that duplex DNA molecules are subject to exonuclease III digestion only after ligation into monomer circles.

By "ligation into monomer circles" is meant ligation of molecules under conditions of dilute concentration such that ends of the same molecule become ligated. Such a procedure is known to those skilled in the art. In these methods it is advantageous sometimes to separate molecules having mismatches from those which do not. By use of appropriate separation procedures both such populations of molecules can be selected.

Methods applying mismatch repair stems to populations of amplified molecules

In another aspect, the invention features a method for correcting base pair mismatches in a population of DNA duplexes that have been produced by enzymatic amplification potentially containing one or more base pair mismatches. The method includes contacting the population of DNA duplexes with a DNA methylase and a mismatch repair system such that base pair mismatches are corrected.

By "enzymatic amplification" is meant a reaction by which DNA molecules are amplified. Examples of such reactions include the polymerase chain reaction and reactions utilizing reverse transcription and subsequent DNA amplification of one or more expressed RNA sequences.

By "mismatch repair system" is meant a complete system such that base pair mismatches are detected and corrected.

In a preferred embodiment, the mismatch repair system is the methyl-directed mismatch repair system of *E. coli*. Components of the defined system capable of correcting mismatches include MutH, MutL, and MutS proteins, DNA helicase II, single-strand binding protein, DNA polymerase III holoenzyme, exonuclease I, exonuclease VII or RecJ, DNA ligase, ATP and four deoxynucleoside triphosphates.

In a further aspect, the invention features a method for removing DNA molecules containing one or more base pair mismatches in a population of molecules that have been produced by enzymatic amplification potentially containing one or more base pair mismatches. The method includes contacting a population of enzymatically amplified molecules with components of a mismatch repair system under conditions such that one or more components of the repair system form a specific complex with a base pair mismatch contained in a DNA duplex and removing DNA duplexes containing the complex from the population of duplex molecules.

By "complex" is meant the result of specific binding of at least one component of mismatch repair system to a base pair mismatch.

26

In a preferred embodiment, the mismatch repair system is the *E. coli* methyl-directed mismatch repair system, the component of the system is the MutS protein, the MutS protein is affixed to a solid support and removal of the DNA duplex containing the complex is by binding to this support.

Methods of attachment of proteins to solid support systems and use of those systems to perform chromatography so as to remove specific molecules are well known to those skilled in the art.

In another embodiment, the invention features a method for removing DNA molecules containing one or more base pair mismatches in a population of DNA duplexes that have been produced by enzymatic amplification, potentially containing one or more base pair mismatches. The method comprises the steps of contacting the population of DNA duplexes with components of a mismatch repair system under conditions such that an endonucleolytic incision is made on a newly synthesized strand of a DNA duplex molecule containing a base pair mismatch so that such a molecule cannot produce a full-sized product in a subsequent round of enzymatic amplification.

By "endonucleolytic cleavage" is meant cleavage on the unmethylated strand at a hemimethylate of GATC sequence by components of a mismatch repair system.

By "full sized product" is meant a molecule that includes the entire region of interest that is subject to amplification. Molecules that contain endonucleolytic cleavage cannot be amplified in subsequent rounds to produce full sized product and thus will be eliminated from the final amplified product population.

In a preferred embodiment the mismatch repair system is the methyl-directed mismatch repair system of *E. coli* and the components are Muts, MutL, and MutH proteins, and ATP.

Methods to remove from a population molecules containing a base pair mismatch

In a further embodiment the invention features a method for removing DNA duplex molecules containing base pair mismatches in a population of heteroduplex DNA molecules produced from different sources. The method comprises contacting the population of DNA duplex molecules potentially containing base pair mismatches with some or all components of a mismatch repair system under conditions such that the component or components form a complex with the DNA having a base pair mismatch, and not with a DNA duplex lacking a base pair mismatch, and removing DNA molecules containing the complex or the product of the complex.

By "product of the complex" is meant a DNA duplex that has incorporated biotinylated nucleotides.

By "some or all components of a mismatch repair system" is meant either a complete mismatch repair system such that the complete reaction is carried out or only the proteins of the system which specifically bind to the mismatch.

In preferred embodiments the mismatch repair system is the methyl-directed mismatch repair system of *E. coli*; some or all protein of the mismatch repair system have been affixed to a solid support and removal by adsorption; the complex interacts with other cellular proteins, and removal of the complex occurs through the interaction; and the conditions include the use of biotinylated nucleotides such that the nucleotides are incorporated into duplex molecules that contained a base pair mismatch and such duplexes are removed by binding to avidin.

By "some or all proteins" is meant, for example, *E. coli* proteins MutS, MutL, and MutH.

By "attached to a solid support" is meant a means, such as by fusion with glutathione transferase, by which a protein is attached to a solid support system and still remains functional.

5,679,522

27

By "adsorption" is meant specific binding to some or all of the proteins of the mismatch repair system affixed to a solid support so that separation from other molecules that do not bind to the solid support affixed proteins occurs.

By "interacts with other cellular proteins" is meant interaction between mismatch repair system protein or between those proteins and other proteins. For example, the interaction of MutS bound to a duplex DNA containing a mismatch with MutL or RecA.

Kits containing a mismatch repair system

In a preferred embodiment, a kit for correcting base pair matches in duplex DNA molecules including one or more of the following components comprising the following purified components: an aliquot of *E. coli* MutH, MutL, and MutS proteins or species variations thereof, an aliquot of DNA helicase II, an aliquot of single-strand DNA binding protein, an aliquot of DNA polymerase III holoenzyme, an aliquot of exonuclease I, an aliquot of Exo VII or RecJ, an aliquot of DNA ligase, an aliquot of ATP, and an aliquot of four deoxynucleoside triphosphates.

A further embodiment of this aspect of this invention includes an assay kit for eliminating DNA molecules containing one or more base pairing mismatches from a population of heterohybrid duplex molecules formed by base pairing of single-stranded DNA molecules obtained from a first and a second source comprising one or more of the following components, an aliquot of proteins of a mismatch repair system, and an aliquot of a recombinase protein.

By "proteins of a mismatch repair system" are meant proteins that modulate the activity of a recombinase protein.

In a preferred embodiment, the proteins of the mismatch correction system are the MutS and MutL proteins of *E. coli*.

Another aspect of the invention features an assay kit for removing DNA molecules containing one or more base pair mismatches comprising an aliquot of one or more proteins of a mismatch repair system that have been affixed to a column support.

In a preferred embodiment, the protein of the mismatch repair system is the MutS protein of *E. coli*.

Another aspect of the invention features a kit for fractionating a heteroduplex DNA population into two pools, one of which was mismatch-free at the beginning of the procedure, the second of which represents duplexes that contained mispaired bases at the beginning of the procedure. This kit is comprised of one or more of the following components: an aliquot of all components of complete mismatch repair system; an aliquot of biotinylated nucleotides; and an aliquot of avidin or an avidin-based support.

In a preferred embodiment, the mismatch repair system is from *E. coli* and consists of products of the mutH, mutL, and mutS genes, DNA helicase II, single-strand DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, exonuclease VII or RecJ exonuclease, DNA ligase, and ATP.

The following Examples are provided for further illustrating various aspects and embodiments of the present invention and are in no way intended to be limiting of the scope.

EXAMPLE 1. DNA Mismatch Correction in a Defined System

In order to address the biochemistry of methyl-directed mismatch correction, the reaction has been assayed in vitro using the type of substrate illustrated in FIG. 1. Application of this method to cell-free extracts of *E. coli* (A. L. Lu, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983) confirmed in vivo findings that methyl-directed repair requires the products of four mutator genes, mutH, mutL,

28

mutS and uvrD (also called mutU), and also demonstrated a requirement for the *E. coli* single-strand DNA binding protein (SSB). The dependence of in vitro correction on mutH, mutL and mutS gene products has permitted isolation of these proteins in near homogeneous, biologically active forms. The MutS protein binds to mismatched DNA base pairs; the MutL protein binds to the MutS-heteroduplex complex (M. Grilley, K. M. Welsh, S. -S. Su, P. Modrich, *J. Biol. Chem.* 264, 1000, 1989); and the 25-kD MutH protein possesses a latent endonuclease that incises the unmethylated strand of a hemimethylated d(GATC) site (K. M. Welsh, A. -L. Lu, S. Clark, P. Modrich, *J. Biol. Chem.* 262, 15624, 1987), with activation of this activity depending on interaction of MutS and MutL with a heteroduplex in the presence of ATP (P. Modrich, *J. Biol. Chem.* 264, 6597, 1989). However, these three Mut proteins together with SSB and the DNA helicase II product of the uvrD (mutU) gene (I. D. Hickson, H. M. Arthur, D. Bramhill, P. T. Emmerson, *Mol. Gen. Genet.* 190, 265, 1983) are not sufficient to mediate methyl-directed repair. Below is described identification of the remaining required components and reconstitution of the reaction in a defined system.

Protein and cofactor requirements for mismatch correction. Methyl-directed mismatch correction occurs by an excision repair reaction in which as much as several kilobases of the unmethylated DNA strand is excised and resynthesized (A. -L. Lin, K. Welsh, S. Clark, S. -S. Su, P. Modrich, *Cold Spring Harbor Symp. Quant. Biol.* 49, 589, 1984). DNA polymerase I, an enzyme that functions in a number of DNA repair pathways, does not contribute in a major way to methyl-directed correction since extracts from a polA deletion strain exhibit normal levels of activity. However extracts derived from a dnaZ⁺ strain are temperature sensitive for methyl-directed repair in vitro (Table 1).

Table 1. Requirement for t and g Subunits of DNA Polymerase III Holoenzyme in Mismatch Repair

TABLE 1			
Requirement for t and g Subunits of DNA Polymerase III Holoenzyme in Mismatch Repair			
Extract genotype	DNA Pol III addition (ng)	Mismatch Correction Activity (fmol/h/mg)	ratio (42°/34°)
Extract preincubation			
42°			
dnaZ ⁺	—	8	910.09
	57 ng	75	1600.47
dnaZ ⁻	—	150	1600.94
	57 ng	160	1601.0

Extracts from strains AX727 (lac thi str^R dnaZ20-16) and AX729 (as AX727 except pure dnaZ⁺) were prepared as described (A. -L. Lin, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983). Samples (110 µg of protein) were mixed with 0.8 µl of 1M KCl and water to yield a volume of 7.2 µl, and preincubated at 42° or 34° C. for 2.5 minutes. All heated samples were then placed at 34° C. and supplemented with 2.2 µl of a solution containing 0.1 µg (24 fmol) of hemimethylated G—T heteroduplex DNA, 16 ng of MutL protein, 50 ng of MutS protein, and buffer and nucleotide components of the mismatch correction assay (A. -L. Lu, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983). DNA polymerase III holoenzyme (57 ng in 0.6 µl) or enzyme buffer was then added, and incubation at 34° C. was continued for 60 min. Heated extracts were supplemented with purified MutL and MutS proteins because these

5,679,522

29

components are labile at 42° C. Activity measurements reflect the correction of heteroduplex sites.

The *dnaZ* gene encodes the τ and γ subunits of DNA polymerase III holoenzyme (M. Kodaira, S. B. Biswas, A. Kornberg, *Mol. Gen. Genet.* 192, 80, 1983; D. A. Mullin, C. L. Woldringh, J. M. Henson, J. R. Walker, *Mol. Gen. Genet.* 192, 73, 1983), and mismatch correction activity is largely restored to heated extracts of the temperature-sensitive mutant strain by addition of purified polymerase III holoenzyme. Since DNA polymerase III holoenzyme is highly processive, incorporating thousands of nucleotides per DNA binding event, the involvement of this activity is consistent with the large repair tracts associated with the methyl-directed reaction.

Additional data indicate that purified MutH, MutL, and MutS proteins, DNA helicase II, SSB, and DNA polymerase III holoenzyme support methyl-directed mismatch correction, but this reaction is inhibited by DNA ligase, an enzyme that is shown below to be required to restore covalent continuity to the repaired strand. This observation led to isolation of a 55-kD stimulatory protein that obviates ligase inhibition. The molecular weight and N-terminal sequence of this protein indicated identity to exonuclease I (G. J. Phillips and S. R. Kushner, *J. Biol. Chem.* 262, 455, 1987), and homogeneous exonuclease I readily substitutes for the 55-kD stimulatory activity (Table 2). Thus, exonuclease I and the six activities mentioned above mediate efficient methyl-directed mismatch correction in the presence of ligase to yield product molecules in which both DNA strands are covalently continuous.

TABLE 2

Stimulation of in vitro Methyl-Directed Correction by Exonuclease I	
Protein added	Mismatch correction (fmol/20 min)
None	1
55-kD protein	18
Exonuclease I	18

Reactions (10 μ l) contained 0.05M HEPES (potassium salt, pH 8.0), 0.02M KCl, 6 mM MgCl₂, bovine serum albumin (0.05 mg/ml), 1 mM dithiothreitol, 2mM ATP, 100 μ M (each) dATP, dCTP, dGTP, and dTTP, 25 μ M β -NAD⁺, 0.1 μ g of hemimethylated, covalently closed G—T heteroduplex DNA (FIG. 1, methylation on c strand, 24 fmol), 0.26 ng of MutH (K. M. Welsh, A. -L. Lin, S. Clark, P. Modrich, *J. Biol. Chem.* 262, 15624, 1987), 17 ng of MutL (M. Grilley, K. R. Welsh, S. -S. Su, P. Modrich, *J. Biol. Chem.* 264, 1000, 1989), 35 ng of MutS (S. -S. Sin and P. Modrich, *Proc. Nat'l Acad. Sci. USA* 83, 5057, 1986), 200 ng of SSB (T. R. Lohman, J. R. Green, R. S. Beyer, *Biochemistry* 25, 21, 1986; U.S. Biochemical Corp.), 10 ng of DNA helicase II (K. Kumura and M. Sekiguchi, *J. Biol. Chem.* 259, 1560, 1984), 20 mg of *E. coli* DNA ligase (U.S. Biochemical Corp.), 95 ng of DNA polymerase III holoenzyme (C. McHenry and A. Kornberg, *J. Biol. Chem.* 252, 6478, 1977), and 1 ng of 55-kD protein or exonuclease I (U.S. Biochemical Corp.) as indicated. Reactions were incubated at 37° C. for 20 minutes, quenched at 55° C. for 10 minutes, chilled on ice, and then digested with Xho I or Hind III endonuclease to monitor correction. Repair of the G—T mismatch yielded a only the G—C containing, Xho I-sensitive product.

The requirements for repair of a covalently closed G—T heteroduplex (FIG. 1) are summarized in Table 3 (Closed

30

circular). No detectable repair was observed in the absence of MutH, MutL, or MutS proteins or in the absence of DNA polymerase III holoenzyme, and omission of SSB or exonuclease I reduced activity by 85 to 90 percent.

TABLE 3

Protein and Cofactor Requirements for Mismatch Correction in a Defined System		
Reaction conditions	Mismatch correction (fmol/20 min)	
	Closed Circular Heteroduplex	Open Circular Heteroduplex
Complete	15	17 (No mutH, No ligase)
minus MutH	<1	—
minus MutL	<1	<1
minus MutS	<1	<1
minus DNA polymerase III holoenzyme	<1	<1
minus SSB	2	1.4
minus exonuclease I	2	<1
minus DNA helicase II	16	15
minus helicase II, plus immune serum	<1	<1
minus helicase II, plus preimmune serum	14	NT
minus Ligase/NAD ⁺	14	NT
minus MgCl ₂	<1	NT
minus ATP	<1	NT
minus dNTP's	<1	NT

Reactions utilizing covalently closed G—T heteroduplex (modification on c strand) were performed as described in the legend to Table 2 except that 1.8 ng of exonuclease I was used. Repair of open circular DNA was performed in a similar manner except that MutH, DNA ligase, and β -NAD⁺ were omitted from all reactions, and the hemimethylated G—T heteroduplex (modification on c strand) had been incised with MutH protein as described in the legend to FIG. 4. When present, rabbit antiserum to helicase II or preimmune serum (5 μ g protein) was incubated at 0° C. for 20 minutes with reaction mixtures lacking MgCl₂; the cofactor was then added and the assay was performed as above. Although not shown, antiserum inhibition was reversed by the subsequent addition of more helicase II. With the exception of the DNA polymerase III preparation, which contained about 15% by weight DNA helicase II (text), the purity of individual protein fractions was \geq 95%. NT—not tested.

These findings are in accord with previous conclusions concerning requirements of the methyl-directed reaction. However, in contrast to observations in vivo and in crude extracts indicating a requirement for the uvrD product, the reconstituted reaction proceeded readily in the absence of the added DNA helicase II (Table 2). Nevertheless, the reaction was abolished by antiserum to homogeneous helicase II, suggesting a requirement for this activity and that it might be present as a contaminant in one of the other proteins. Analysis of these preparations for their ability to restore mismatch repair to an extract derived from a uvrD (*mutI*) mutant and for the physical presence of helicase II by immunoblot assay revealed that the DNA polymerase III holoenzyme preparation contained sufficient helicase II (13 to 15 percent of total protein by weight) to account for the levels of mismatch correction observed in the defined system. Similar results were obtained with holoenzyme preparations obtained from two other laboratories. The purified system therefore requires all the proteins that have been previously implicated in methyl-directed repair.

5,679,522

31

The rate of correction of the closed circular heteroduplex was unaffected by omission of DNA ligase (Table 3), but the presence of this activity results in production of a covalently closed product. Incubation of a hemimethylated, supercoiled G—T heteroduplex with all seven proteins required for correction in the presence of DNA ligase resulted in extensive formation of covalently closed, relaxed, circular molecules. Production of the relaxed DNA was dependent on MutS (FIG. 2) and MutL proteins, and the generation of this species was associated with heteroduplex repair (FIG. 2). Correction also occurred in the absence of ligase, but in this case repair products were open circular molecules, the formation of which depended on the presence of MutS (FIG. 2). Since MutS has no known endonuclease activity but does recognize mispairs, it is inferred that open circular molecules are the immediate product of a mismatch-provoked-excision repair process. Ligase closure of the strand break(s) present in this species would yield the covalently closed, relaxed circular product observed with the complete system.

The set of purified activities identified here as being important in methyl-directed repair support efficient correction. In the experiments summarized in Table 3, the individual proteins were used at the concentrations estimated to be present in the standard crude extract assay for correction as calculated from known specific activity determinations. Under such conditions the rate and extent of mismatch repair in the purified system are essentially identical to those observed in cell-free extracts.

DNA sites involved in repair by the purified system. The single d(GATC) sequence within the G—T heteroduplex shown in FIG. 1 is located 1024 base pairs from the mismatch. Despite the distance separating these two sites, correction of the mismatch by the purified system responded to the state of modification of the d(GATC) sequence as well as its presence within the heteroduplex (FIG. 3). A substrate bearing d(GATC) methylation on both DNA strands did not support mismatch repair nor did a related heteroduplex in which the d(GATC) sequence was replaced by d (GATT). However, each of the two hemimethylated heteroduplexes were subject to strand-specific correction, with repair in each case being restricted to the unmodified DNA strand. With a heteroduplex in which neither strand was methylated, some molecules were corrected on one strand, and some were corrected on the other. As can be seen, the hemimethylated heteroduplex bearing methylation on the complementary DNA strand was a better substrate than the alternative configuration in which modification was on the viral strand, with a similar preference for repair of the viral strand being evident with the substrate that was unmethylated on either strand. This set of responses of the purified system to the presence and state of modification of d(GATC) sites reproduce effects previously documented in vivo and in crude extract experiments (R. S. Lahue, S. -S. Su, P. Modrich, *Proc. Natl. Acad. Sci. USA* 84, 1482, 1987).

TABLE 4

Correction Efficiencies for Different Mismatches.					
Heteroduplex	Markers	C ⁺ V ⁻		C ⁻ V ⁺	
		Rate	Bias	Rate	Bias
C 5'-CTCGA G AGCTT	Xho I	1.2	>18	0.38	>5
V 3'-GAGCT T TCGAA	Hind III				
C 5'-CTCGA G AGCTG	Xho I	1.1	>17	0.38	>6
V 3'-GAGCT G TCGAC	Pvu II				
C 5'-ATCGA T AGCTT	Cla I	1.0	>16	0.24	3

32

TABLE 4-continued

Correction Efficiencies for Different Mismatches.					
Heteroduplex	Markers	C ⁺ V ⁻		C ⁻ V ⁺	
		Rate	Bias	Rate	Bias
V 3'-TAGCT T TCGAA	Hind III				
C 5'-ATCGA A AGCTT	Hind III	0.88	>20	0.20	>7
V 3'-TAGCT A TCGAA	Cla I				
C 5'-CTCGA A AGCTT	Hind III	0.61	17	0.28	>5
V 3'-GAGCT C TCGAA	Xho I				
C 5'-GTCTGA C AGCTT	Sal I	0.60	12	0.23	>4
V 3'-CAGCT T TCGAA	Hind III				
C 5'-GTCTGA A AGCTT	Hind III	0.44	>13	0.21	5
V 3'-CAGCT T TCGAA	Sal I				
C 5'-CTCTGA C AGCTG	Pvu II	0.04	NS	<0.04	NS
V 3'-GAGCT C TCGAC	Xho I				

Table 4. (Continued) Correction of the eight possible base-base mispairs was tested with the set of covalently closed heteroduplexes described previously including the G—T substrate shown in FIG. 1. With the exception of the mispair and the variations shown at the fifth position on either side, all heteroduplexes were identical in sequence. Each DNA was tested in both hemimethylated configurations under complete reaction conditions (Table 3, closed circular heteroduplex) except that samples were removed at 5-minute intervals over a 20 minute period in order to obtain initial rates (fmol/min). c and v refer to complementary and viral DNA strands, and Bias indicates the relative efficiency of mismatch repair occurring on the two DNA strands (ratio of unmethylated to methylated) as determined 60 minutes after the reaction was started. NS—not significant. With the exception of the C—C heteroduplexes, repair in the absence of MutS protein was less than 20% (in most cases <10%) of that observed in its presence (not shown).

The efficiency of repair by the methyl-directed pathway depends not only on the nature of the mispair, but also on the sequence environment in which the mismatch is embedded (P. Modrich, *Ann. Rev. Biochem.* 56, 435, 1987). To assess the mismatch specificity of the purified system under conditions where sequence effects are minimized, a set of heteroduplexes were used in which the location and immediate sequence environment of each mispair are essentially identical (S. -S. Su, R. S. Lahue, K. G. Au, P. Modrich, *J. Biol. Chem.* 263, 6829, 1988). This analysis (Table 4) showed that the purified system is able to recognize and repair in a methyl-directed manner seven of the eight possible base-base mismatches, with C-C being the only mispair that was not subject to significant correction. Table 3 also shows that the seven corrected mismatches were not repaired with equal efficiency and that in the case of each heteroduplex, the hemimethylated configuration modified on the complementary DNA strand was a better substrate than the other configuration in which the methyl group was on the viral strand. These findings are in good agreement with patterns of repair observed with this set of heteroduplexes in *E. coli* extracts (Although the patterns of substrate activity observed in extracts and in the purified system are qualitatively identical, the magnitude of variation observed differs for the two systems. Hemimethylated heteroduplexes modified on the complementary DNA strand are better substrates in both systems, but in extracts such molecules are repaired at about twice the rate of molecules methylated on the viral strand. In the purified system these relative rates differ by factors of 2 to 4. A similar effect may also exist with respect to mismatch preference within a given hemimethylated family. Although neither system repairs C-C, the rates

5,679,522

33

of repair of other mismatches vary by a factors of 1.5 to 2 in extracts but by factors of 2 to 3 in the defined system.).

Strand-specific repair directed by a DNA strand break. Early experiments on methyl-directed repair in *E. coli* extracts led to the proposal that the strand-specificity of the reaction resulted from endonucleolytic incision of an unmethylated DNA strand at a d(GATC) sequence. This idea was supported by the finding that purified MutH protein has an associated, but extremely weak d(GATC) endonuclease that is activated in a mismatch-dependent manner in a reaction requiring MutL, MutS, and ATP. The purified system has been used to explore this effect more completely.

The two hemimethylated forms of the G—T heteroduplex shown in FIG. 1 were incised using high concentrations of purified MutH protein to cleave the unmethylated DNA strand at the d(GATC) sequence (>pGpApTpC). After removal of the protein, these open circular heteroduplexes were tested as substrates for the purified system in the absence of DNA ligase. Both open circular species were corrected in a strand-specific manner and at rates similar to those for the corresponding covalently closed heteroduplexes (FIG. 4). As observed with closed circular heteroduplexes, repair of the MutH-cleaved molecules required MutL, MutS, SSB, DNA polymerase III holoenzyme, and DNA helicase II (FIG. 4 and open circle entries of Table 2), but in contrast to the behavior of the closed circular substrates, repair of the mismatch within the open circular molecules occurred readily in the absence of MutH protein. Thus prior incision of the unmethylated strand of a d(GATC) site can bypass the requirement for MutH protein in strand-specific mismatch correction.

The nature of the MutH-independent repair was examined further to assess the effect of ligase on the reaction and to determine whether a strand break at a sequence other than d(GATC) can direct correction in the absence of MutH protein (FIG. 5). As mentioned above, a covalently closed G—T heteroduplex that lacks a d(GATC) sequence is not subject to repair by the purified system in the presence (FIG. 3) or absence of DNA ligase. However, the presence of one strand-specific, site-specific break is sufficient to render this heteroduplex a substrate for the purified system in the absence of ligase and Ruth protein (FIG. 5). Repair of this open circular heteroduplex was limited to the incised, complementary DNA strand, required presence of MutL and MutS proteins, DNA polymerase III, and SSB, and correction of the molecule was as efficient as that observed with the hemimethylated heteroduplex that had been cleaved by MutH at the d(GATC) sequence within the complementary strand. Although the presence of a strand break is sufficient to permit strand-specific correction of a heteroduplex in the absence of MutH and ligase, the presence of the latter activity inhibited repair not only on the heteroduplex lacking a d(GATC) sequence but also on both hemimethylated molecules that had been previously incised with MutH protein (FIG. 5). This inhibition by ligase was circumvented by the presence of MutH protein, but only if the substrate contained a d(GATC) sequence, with this effect being demonstrable when both types of heteroduplex were present in the same reaction (FIG. 5, last column). This finding proves that MutH protein recognizes d(GATC) sites and is consistent with the view that the function of this protein in mismatch correction is the incision of the unmethylated strand at this sequence.

EXAMPLE 2: Purification of MutY Protein

Purification of MutY Protein *E. coli* RK1517 was grown at 37° C. in 170 liters of L broth containing 2.5 mM

34

KH_2PO_4 , 7.5 mM Na_2HPO_4 (culture, pH=7.4) and 1% glucose. The culture was grown to an A_{590} of 4, chilled to 10° C. and cells were harvested by continuous flow centrifugation. Cell paste was stored at 70° C. A summary of the MutY purification is presented in Table 1. Fractionation procedures were performed at 0°–4° C., centrifugation was at 13,000×g, and glycerol concentrations are expressed as volume percent.

Frozen cell paste (290 g) was thawed at 4° C., resuspended in 900 ml of 0.05 M Tris-HCl (pH 7.5), 0.1M NaCl, 1 mM dithiothreitol, 0.1 mM EDTA, and cells were disrupted by sonication. After clarification by centrifugation for 1 hr, the lysate (Fraction I, 970 ml) was treated with 185 ml of 25% streptomycin sulfate (wt/vol in 0.05M Tris-HCl (pH 7.5), 0.1M NaCl, 1 mM dithiothreitol, 0.1 mM EDTA) which was added slowly with stirring. After 30 min of additional stirring, the solution was centrifuged for 1 h, and the supernatant (1120 ml) was treated with 252 g of solid ammonium sulfate which was added slowly with stirring. After 30 min. of additional stirring, the precipitate was collected by centrifugation for 1 h, resuspended to a final volume of 41 ml in 0.02M potassium phosphate (pH 7.5), 0.1 mM EDTA, 10% (vol/vol) glycerol, 1 mM dithiothreitol, and dialyzed against two 2 l portions of 0.02M potassium phosphate (pH 7.5), 0.1M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol (2 h per change). The dialyzed material was clarified by centrifugation for 10 min to yield Fraction II (45 ml).

Fraction II was diluted 10-fold into 0.02M potassium phosphate (pH 7.5), 0.1M EDTA, 1 mM dithiothreitol, 10% glycerol so that the conductivity of the diluted solution was comparable to that of the dilution buffer containing 0.1M KCl. The solution was performed on small aliquots of Fraction II, and diluted samples were immediately loaded at 1 ml/min onto a 14.7 cm×12.6 cm² phosphocellulose column equilibrated with 0.02M potassium phosphate (pH 7.5), 0.1M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. The column was washed with 400 ml of equilibration buffer, and developed with a 2 liter linear gradient of KCl (0.1 to 1.0M) in 0.02M potassium phosphate (pH 7.5), 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. Fractions containing MutY activity, which eluted at about 0.4M KCl, were pooled (Fraction III, 169 ml).

Fraction III was dialyzed against two 500 ml portions of 5 mM potassium phosphate (pH 7.5), 0.5M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol (2 h per change) until the conductivity was comparable to that of the dialysis buffer. After clarification by centrifugation at for 10 min, the solution was loaded at 0.5 ml/min onto a 21 cm×2.84 cm² hydroxylapatite column equilibrated with 5 mM potassium phosphate, pH 7.5, 0.05M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. After washing with 130 ml of equilibration buffer, the column was eluted with a 600 ml linear gradient of potassium phosphate (5 mM to 0.4M, pH 7.5) containing 0.05M KCl, 1 mM dithiothreitol, 10% glycerol. Fractions eluting from the column were supplemented with EDTA to 0.1 mM. Peak fractions containing 60% of the total recovered activity, which eluted at about 0.1M potassium phosphate, were pooled (Fraction IV, 24 ml). The remaining side fractions contained impurities which could not be resolved from MutY by MonoS chromatography.

Fraction IV was diluted by addition of an equal volume of 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. After clarification by centrifugation for 15 min, diluted Fraction IV was loaded at 0.75 ml/min onto a Pharmacia HR 5/5 MonoS FPLC column that was equilibrated with 0.05M sodium

5,679,522

35

phosphate (pH 7.5), 0.1M NaCl, 0.1mM EDTA, 0.5 mM dithiothreitol, 10% glycerol. The column was washed at 0.5 ml/min with 17 ml of equilibration buffer.

Ex 2/Table 5

TABLE 5

Purification of MutY protein from 290 g of <i>E. coli</i> RK1517				
Fraction	Step	Total Protein mg	Specific Activity units/mg	Yield Percent
I	Extract	10,900	40	(100)
II	Ammonium sulfate	1,350	272	84
III	Phosphocellulose	66	10,800	160
IV	Hydroxylapatite	1.4	136,000	44
V	MonoS	0.16	480,000	18

Specific A*G to C—G mismatch correction in cell-free extracts was determined as described previously (Au et al. 1988), except that ATP and glutathione were omitted from the reaction and incubation was for 30 min instead of 1 h. For complementation assays, each 0.01 ml reaction contained RK1517-Y33 extract (mutS mutY) at a concentration of 10 mg/ml protein. One unit of MutY activity is defined as the amount required to convert 1 fmol of A*G mismatch to C—G base pair per h under complementation conditions.

20 ml linear gradient of NaCl (0.1 to 0.4M) in 0.05M sodium phosphate (pH 7.5), 0.1 mM EDTA, 0.5 mM dithiothreitol, 10% glycerol. Fractions with MutY activity, which eluted at approximately 0.2M NaCl, were pooled (Fraction V, 2.6 ml). Fraction V was divided into small aliquots and stored at -70° C.

Assay for MutY-dependent, A*G-specific glycosylase

DNA restriction fragments were labeled at either the 3' or 5' ends with ³²P. Glycosylase activity was then determined in 0.01 ml reactions containing 10 ng end-labeled DNA fragments, 0.02M Tris-HCl, pH7.6, 1 mM EDTA, 0.05 mg/ml bovine serum albumin, and 2.7 ng MutY. After incubation at 37° C. for 30 min, the reaction mixture was treated with 2.5×10⁻³ units of HeLa AP endonuclease II in the presence of 11 mM MgCl₂ and 0.005% Triton X-100 for 10 min at 37° C. Reactions were quenched by the addition of an equal volume of 80% formamide, 0.025% xylene cyanol, 0.025% bromphenol blue, heated to 80° C. for 2 min, and the products analyzed on an 8% sequencing gel. Control reactions contained either no MutY, no A*G mismatch or no AP endonuclease II.

Strand cleavage at the AP site generated by MutY could also be accomplished by treatment with piperidine instead of treatment with AP endonuclease II. After incubation for 30 min. at 37° C. with MutY as described above, the reaction mixture was precipitated with ethanol in the presence of carrier tRNA, then resuspended in 1M piperidine and heated at 90° C. for 30 min. After two additional ethanol precipitations, changing tubes each time, the pellet was resuspended in a minimum volume of water to which was added an equal volume of 80% formamide, 0.025% xylene cyanol, 0.025% bromphenol blue. The products were then analyzed on an 8% sequencing gel.

EXAMPLE 3: Genetic Mapping Point Mutations in the Human Genome

The full novelty and utility of the present invention may be further appreciated by reference to the following brief description of selected specific embodiments which advantageously employ various preferred forms of the invention

36

as applied to a common problem in genetic mapping of point mutations in the human genome. In the course of constructing gene linkage maps, for example, it is frequently desirable to compare the sequence of a cloned DNA fragment with homologous sequences in DNA extracted from a human tissue sample. Substantially all base pairs in the entire homologous sequence of the cloned DNA fragment are compared to those of the human tissue DNA, most advantageously in a single test according to the present invention, merely by contacting both strands of the human tissue DNA molecule with both radiolabeled complementary strands of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting DNA duplexes with the *E. coli* MutS protein that recognizes substantially all base pair mismatches under conditions such that the protein forms specific complexes with its cognate mispairs, and detecting the resulting DNA:protein complexes by contacting the complexes with a membranous nitrocellulose filter under conditions such that protein:DNA complexes are retained while DNA not complexed with protein is not retained, and measuring the amount of DNA in the retained complexes by a standard radiological method or by utilizing any of the other methods of the invention; e.g., altered electrophoretic mobility, or detection by use of antibodies.

If the above detection test indicates the presence of sequence differences between the human tissue DNA and the cloned DNA and localization is required, or, in the alternative, if such differences are suspected and localization as well as detection of them is desired in a first analysis, the another method of this invention may be applied for these purposes. An embodiment of this aspect of the invention that may be most advantageously employed comprises the steps of contacting both strands of the human tissue DNA molecule with both radiolabeled complementary strands of the second DNA molecule (usually without separation from the cloning vector DNA) under conditions such that base pairing occurs, contacting the resulting DNA duplexes with MutHLS to produce a GATC cleavage reaction or a modified form of MutS protein of *E. coli* to which is attached an hydroxyl radical cleaving function under conditions such that the radical cleaving function cleaves both strands of the DNA within about 20 base pairs of substantially all DNA base mispairs. In the absence of any DNA base mispairs in the DNA duplexes comprising complementary strands of the human tissue and cloned DNAs, no DNA fragments smaller than the cloned DNA (plus vector DNA, if still attached) would be detected. Determination of the location of any double-stranded DNA cleavages by the modified MutS protein to within a few kbp or less of some restriction enzyme cleavage site within the cloned DNA is determined by standard restriction enzyme mapping approaches. If greater precision in localization and identification of a single base difference is desired, sequencing could be confined to those particular fragments of cloned DNA that span at least one base sequence difference localized by this method and are cleaved by a restriction enzyme at the most convenient distance of those sequence differences for direct sequencing.

The examples herein can be changed to make use of other methods of separation to identify mismatches, such as a filter-binding assay, as well as the nicking reaction with MutS and MutL. While large (at least 20 kbp) or small DNA molecules can be used in these methods those of between 1-10 kbp are preferred.

EXAMPLE 4: DNA Mismatch Detection Kit

Kit contains MutS protein, dilution buffer, annealing buffer, reagents to generate complementary and mismatched

5,679,522

37

control duplexes and filter binding protocol. It can be used to detect single-base mismatches in oligonucleotides.

MutS kit components:

MutS protein in storage buffer: 50mM HEPES pH7.2, 100 mM KCl, 1 mM EDTA, 1 mM DTT;

MutS1: 16mer oligonucleotide GATCCGTCGACCT-GCA (all such oligonucleotides are written 5' to 3' herein) in water (2 μ M);

MutS2: 16mer oligonucleotide TGCAGGTCGACG-GATC 1 μ M in annealing buffer 1 μ M: 20 mM Tris/HCl pH 7.6, 5 mM, MgCl₂, 0.1 mM DTT, 0.01 mM EDTA;

MutS3: 16mer oligonucleotide TGCAGGTTGACG-GATC 1 μ M in annealing buffer;

Assay buffer/annealing buffer/wash buffer, 20 mM Tris/HCl pH 7.6, 5 mM MgCl₂, 0.1 mM DTT, 0.01 mM EDTA;

Protein storage/dilution buffer: 50 mM HEPES pH 7.2, 100 mM KCl, 1 mM EDTA, 1 mM DTT.

The DNA mismatch detection kit contains three 16-mer oligonucleotides labeled MUTS1, MUTS2, and MUTS3 for testing the performance of MutS protein. When MUTS1 and MUTS2 are annealed, a perfectly matched duplex results. When MUTS1 and MUTS3 are annealed, a duplex containing a single G—T mismatch results. These serve as control substrates for MutS binding.

Kinase Labeling of MUTS1 Oligonucleotide

This protocol uses half the amount of oligonucleotide contained in the kit. To a microcentrifuge tube on ice add the following:

MUTS1 Oligonucleotide (2 μ M)	15 μ l (30 pmoles)
10X T4 Polynucleotide Kinase Buffer	3 μ l
³² P-ATP (3000Ci/mmol)	1 μ l
ATP (10 μ M)	2.5 μ l
Sterile dH ₂ O	7.5 μ l
T4 Polynucleotide Kinase (30 units/ μ l)	1 μ l (30 units)

Incubate the reaction mixture for 10 min at 37° C. Then incubate 10 min at 70° C. Spot two independent 1 μ l aliquots of the mixture on a SureCheck TLC plate and also spot a dilution of ³²P-ATP (1:30 in water) in a separate lane and run with the elution mixture. Expose the developed plate to X-ray film for 5 min. Scrape all radioactive spots from both experimental lanes of the plate and count them in a liquid scintillation counter to determine the % incorporation of label. This value is typically 40–60%. If a significant labeled ATP spot is present in the kinase reaction lanes on the plate, the labeled oligonucleotide must be purified before use (TLC or gel), since ³²P-ATP will contribute to background in the filter binding assay. In our experience, this is usually not necessary.

Keep in mind that the MUTS1 oligo stock is 2 pmol/ μ l and that the final concentration should be 1 pmol/ μ l. It is critical that this final concentration be as exact as possible, since the concentration determines the amount of MUTS1 in the next (annealing) step and hence, the amount of DNA available for binding by the protein.

Annealing Reactions

Two separate reactions are carried out: MUTS1/MUTS2 and MUTS1/MUTS3. In both cases, the ³²P-labeled MUTS1 from Step 1 is used.

38

Complementary		Mismatched	
MUTS1 (kinased)	14 μ l = 14 pmol	MUTS1 (kinased)	14 μ l = 14 pmols
MUTS2 (1 μ M)	28 μ l = 28 pmol	MUTS3 (1 μ M)	28 μ l = 28 pmols
annealing buffer	28 μ l	annealing	28 μ l
		buffer	
	70 μ l		70 μ l

1. Heat each mixture for 10 min at 70° C.
2. Incubate for 30 min at room temperature.
3. Hold on ice until ready to use.

The molar ratio of MUTS2/MUTS1 and MUTS3/MUTS1 is 2:1 in the above reactions and this should be maintained for optimal results. Lowering the ratio of unlabeled to labeled strand may lead to very high background in the filter binding assay, presumably caused by sticking of labeled ssDNA to nitrocellulose.

Assay of MutS Binding by the Gel Shift Method

The binding of MutS to mismatches can be assessed using the technique of Gel Shift Mobility Assay (GSMA), a useful tool to identify protein-DNA interactions which may regulate gene expression. Below is a protocol for performing GSMA on the MUTS1/MUTS3 mismatched duplex contained in the mismatch detection kit. Optimum conditions may vary depending on the particular mismatch being detected or the length of the oligonucleotide.

All binding reactions should be carried out on ice. The total binding reaction volume is 10 μ l. Add 4 μ l of a MutS protein dilution (prepared using dilution buffer in the kit) containing 0.5–5 pmoles (0.125–1.25 units) of MutS protein (1 pmol=97 ng) to 6 μ l=1.2 pmoles of ³²P-labeled MUTS1/MUTS3 heteroduplex. Also add comparable amounts of MutS protein to labeled MUTS1/MUTS2 matched duplex to serve as a control. A control incubation consisting only of mismatched heteroduplex (no MutS protein) should also be run. Incubate all reactions on ice for 30 min.

To 3 μ l of the DNA/MutS mixture from each incubation add 1 μ l of a 50% w/v sucrose solution.

Load 2 μ l of the mixture from Step 2 onto a 6% non-denaturing polyacrylamide gel prepared in Tris-acetate-EDTA (TAE) buffer (Sambrook et al., "Molecular Cloning": A Laboratory Manual, Second Edition, Cold Spring Harbor Laboratory, New York (1989)) to which MgCl₂ has been added to a final concentration of 1 mM and run the gel at 10 V/cm and 4° C. in TAE buffer containing 1 mM MgCl₂ until bromophenol blue dye (loaded into an adjacent well) has migrated approximately half way down the gel. The presence of Mg++ in the gel and running buffer is critical for optimal results in the GSMA assay of MutS protein.

Filter Binding Assay

The total binding reaction volume is 10 μ l. It consists of 6 μ l, or 1.2 pmoles, of duplex DNA and 4 μ l of a MutS protein dilution containing 0.5–5 pmoles (0.125–1.25 units) of MutS protein (1 pmol=97 ng). Each type of duplex, complementary and mismatched, should be assayed in duplicate or triplicate along with a no protein control for each annealing, which will serve as the background to subtract.

In order to use the filter binding assay it will be necessary to make up additional annealing buffer for use in the washing step. Add 20 ml of 1M Tris-HCl, pH 7.6, 5 ml of 1M MgCl₂, 0.1 ml of 1M DTT, and 0.02 ml of 0.5M EDTA to distilled water and bring the volume to 1 liter.

For each binding assay, add the following to a 0.5 ml microcentrifuge tube on ice:

5,679,522

39

MUTS1/MUTS2 (Control) OR

MUTS1/MUTS3 (Mismatched)

Annealing Mixture 6 μ l

Set up the filtration apparatus and presoak the nitrocellulose filters in annealing buffer.

Add 4 μ l of MutS protein dilution to the annealing mixtures on ice. Also include no protein controls for each annealing.

After 30 minutes, begin filtration of samples. Caution, use a slow rate of filtration. It should take at least a second or two for the 10 μ l sample to filter.

Immediately wash the filters with 5 ml each of cold annealing buffer. This should take 20–30 seconds.

Place the filters in liquid scintillation vials, add fluid and count for 2 minutes each.

Determine the input cpm for each annealing as follows: To 6 μ l of annealing mixture, add 54 μ l of water and count 2–3 aliquots of 6 μ l each in scintillation fluid. The input cpm is then 10 \times the average of the cpm of the dilution.

Determine the cpm/pmol of DNA as follows:

$$\frac{\text{cpm of 6 } \mu\text{l aliquot} \times \text{dilution} \times \text{fraction of label incorporate}}{\text{pmol of DNA in annealing reaction}}$$

A 6 μ l annealing contains 1.2 pmoles of DNA

A typical kinase reaction may give 42% incorporation (determined previously)

A 6 μ l aliquot of 10 \times dilution may be 10,600 cpm

$$\frac{10,600 \times 10 \times 0.42}{1.2} = 37,100 \text{ cpm/pmol DNA}$$

Determine the pmoles of DNA bound by various pmoles of MutS. First, determine the pmoles of MutS protein in a binding reaction:

$$\frac{\text{concentration of MutS} \times \text{volume of protein added}}{\text{molecular weight of MutS} \times \text{dilution factor}}$$

Example: If 4 μ l of a 6 \times dilution of MutS at 250 μ g/ml is used, then:

$$\frac{250 \text{ ng}/\mu\text{l} \times 4 \mu\text{l}}{97 \text{ ng/pmol} \times 6} = 1.72 \text{ pmoles of MutS in reaction}$$

Then, determine the pmoles of DNA bound:

$$\frac{\text{cpm retained on filter with MutS protein} - \text{cpm on no protein filter}}{\text{cpm/pmol of DNA}}$$

Example: One gets 15,470 cpm on the filter with MutS and 340 cpm with no protein

$$\frac{15,470 \text{ cpm} - 340 \text{ cpm}}{37,100 \text{ cpm/pmol}} = 0.408 \text{ pmoles of DNA bound}$$

Determine the number of pmoles of MutS required to bind 1 pmole of DNA (i.e., a unit of MutS).

In the above example, 1.72 pmoles of MutS bound 0.409 pmoles of DNA, such that one unit = 1.72/0.408 = pmoles MutS per mole DNA.

EXAMPLE 5: Effects of MutS and MutL on RecA-catalyzed Strand Transfer

A model system used to evaluate MutS and MutL effects on RecA catalyzed strand transfer is depicted in FIG. 6. The

40

assay for RecA-catalyzed strand transfer between homologous and quasi-homologous DNA sequences employed the three strand reaction in which one strand from a linear duplex DNA is transferred to an homologous, single-stranded DNA circle (Cox, 78 *Proc. Natl. Acad. Sci. USA* 3433, 1981. These experiments exploited the previous observation that RecA is able to support strand transfer between related fd and M13 DNAs (Bianchi et al., 35 *Cell* 511, 1983; DasGupta et al., 79 *Proc. Natl. Acad. Sci. USA* 762, 1982, which are approximately 97% homologous at the nucleotide level. The vast majority of this variation is due to single base pair changes.

Results of experiments on the effects of MutS and MutL on RecA-catalyzed strand transfer between homologous and quasi-homologous DNA sequences are shown in FIG. 7. Reactions (50 μ l) contained 50 mM HEPES (pH 7.5), 12 mM MgCl_2 , 2 mM ATP, 0.4 mM dithiothreitol, 6 mM phosphocreatine, 10 U/ml phosphocreatine kinase, 0.6 nM single-stranded circular DNA (molecules), 7.6 μ g RecA protein, 0.54 μ g SSB, and MutS or MutL as indicated. Reactions were allowed to preincubate at 37° C. for 10 minutes, strand exchange was initiated by addition of linear duplex fd DNA (Rf DNA linearized by cleavage with HpaI, 0.6 nM final concentration as molecules), and incubation continued for 70 minutes. MutS or MutL was added 1 minute prior to addition of duplex DNA. Sample (50 μ l) were quenched by addition of EDTA (25 mM), sodium dodecyl sulphate (0.1%), and proteinase K (150 μ g/ml), followed by incubations at 42° C. for 30 minutes.

The presence of MutS or MutL was without significant effect on strand transfer between linear duplex fd DNA and circular fd single-strands, MutS did inhibit strand transfer between quasi-homologous linear duplex fd DNA and M13 single-strands. Similar results were obtained for strand transfer between duplex M13 DNA and single-stranded fd (data not shown). In contrast, MutL alone did not significantly alter the yield of circular duplex product formed by RecA catalyzed strand transfer between these different DNAs.

EXAMPLE 6: MutL Potentiation of MutS Block to Strand Transfer

Results of experiments on the MutL potentiation of the MutS block to strand transfer in response to mismatched base pairs are shown in FIG. 8. Reaction mixtures (210 μ l) contained 50 mM HEPES (pH 7.5), 12 mM MgCl_2 , 2 mM ATP, 0.4 mM dithiothreitol, 6 mM phosphocreatine, 10 U/ml phosphocreatine kinase, 0.6 nM (molecules) single-stranded circular DNA, 32 μ g RecA protein, and 2.3 μ g SSB. Reactions were preincubated for 10 minutes at 37° C. and strand exchange initiated by addition of duplex fd DNA (Rf DNA linearized by cleavage with HpaI, 0.6 nM final concentration as molecules). When present, MutS (2.9 μ g) and/or MutL (1.3 μ g) were added 1 minute prior to addition of duplex DNA. Samples were removed as indicated times and quenched as described in Example 5.

MutL potentiates the inhibition of heteroduplex formation that is observed with MutS. Formation of full length, circular heteroduplex product is virtually abolished in the presence of MutS and MutL. Heteroduplex formation between perfectly homologous strands occurred readily in the presence of either or both proteins.

EXAMPLE 7: MutS and MutL Block of Branch Migration

While MutS and MutS along with MutL blocked formation of fully duplex, circular fd-M13 product, some strand

5,679,522

41

transfer did occur in these reactions as demonstrated by the occurrence of strand transfer "intermediates" that migrated more slowly in agarose gels than fully duplex, nicked circular product (data not shown). The nature of these structures was examined using the S1 nuclease procedure of Cox and Lehman to evaluate mean length of stable heteroduplex formation. This analysis is shown in FIG. 9.

Reaction mixtures (510 μ l) contained 50 mM HEPES (pH 7.5), 12 mM $MgCl_2$, 2 mM ATP, 0.4 mM dithiothreitol, 6 mM phosphocreatine, 10 U/mL phosphocreatine kinase, 0.6 nM single-stranded circular DNA (molecules), 77 μ g RecA protein, 5.5 μ g SSB, and when indicated 6.9 μ g MutS and 3.2 μ g MutL. Reactions were allowed to preincubate at 37° C. for 10 minutes, strand exchange was initiated by addition of linear duplex [3H]M13 DNA (Rf DNA linearized by cleavage with HpaI, 0.6 nM final concentration as molecules). MutS or MutL was added 1 minute prior to addition of M13 duplex DNA. Samples (100 μ l) were taken as indicated, quenched with sodium dodecyl sulphate (0.8%), and extracted with phenol:chloroform:isoamyl alcohol (24:24:1) equilibrated with 10 mM Tris-HCl, pH 8.0, 0.1 mM EDTA. The organic phase was back-extracted with 0.5 volume of 50 mM HEPES, pH 5.5. Aqueous layers were combined washed with H_2O -saturated ether, and relieved of residual ether by 30 minutes incubation at 37° C. The mean length of stable heteroduplex was then determined using S1 nuclease (10 U/ml) according to Cox and Lehman (Cox, 1981 supra).

Although some strand transfer occurs between fd and M13 DNAs in the presence of MutS and MutL, heteroduplex formation is restricted to about one kilobase of the 6.4 kilobase possible. The MutS and MutL effects on recombination are due, at least in part, to their ability to control branch migration reaction in response to occurrence of mismatched base pairs.

Other embodiments are within the following claims.

What is claimed is:

1. A method for detecting a base pair mismatch in a DNA duplex, comprising the steps of:

contacting at least one strand of a first DNA molecule with the complimentary strand of a second DNA molecule under conditions such that base pairing occurs;

contacting a DNA duplex potentially containing a base pair mismatch with a mispair recognition protein under conditions suitable for said protein to form a specific complex only with said DNA duplex having a base pair mismatch, and not with a DNA duplex with conventional Watson-Crick base pairing,

detecting any DNA:protein complexes as a measure of the presence of a base pair mismatch in said DNA duplex by

contacting said DNA:protein complexes with a selectively adsorbent agent under conditions such that said DNA:protein complexes are retained on said agent while DNA not complexed with protein is not retained, and

measuring the amount of DNA in said retained complexes.

2. The method of claim 1 wherein said adsorbent agent is a membranous nitrocellulose filter.

3. A method for detecting a base pair mismatch in a DNA duplex, comprising the steps of:

contacting at least one strand of a first DNA molecule with the complimentary strand of a second DNA molecule under conditions such that base pairing occurs;

contacting a DNA duplex potentially containing a base pair mismatch with a mispair recognition protein under

42

conditions suitable for said protein to form a specific complex only with said DNA duplex having a base pair mismatch, and not with a DNA duplex with conventional Watson-Crick base pairing,

detecting any DNA:protein complexes as a measure of the presence of a base pair mismatch in said DNA duplex by comparing the electrophoretic mobility of said DNA:protein complexes to that of uncomplexed DNA.

4. A method for detecting a base pair mismatch in a DNA duplex, comprising the steps of:

contacting at least one strand of a first DNA molecule with the complimentary strand of a second DNA molecule under conditions such that base pairing occurs;

contacting a DNA duplex potentially containing a base pair mismatch with a mispair recognition protein wherein said mispair recognition protein is the product of the mutS gene of *Salmonella typhimurium*, the hexA gene of *Salmonella pneumoniae* or the Msh1; Msh2, genes of yeast, under conditions suitable for said protein to form a specific complex only with said DNA duplex having a base pair mismatch and not with a DNA duplex with conventional Watson-Crick base pairing, and

detecting any said complex as a measure of the presence of a base pair mismatch in said DNA duplex.

5. A method for detecting a base pair mismatch in a DNA duplex, comprising the steps of:

contacting at least one strand of a first DNA molecule with the complimentary strand of a second DNA molecule under conditions such that base pairing occurs;

contacting a DNA duplex potentially containing a base pair mismatch with a mispair recognition protein under conditions suitable for said protein to form a specific complex only with said DNA duplex having a base pair mismatch, and not with a DNA duplex with conventional Watson-Crick base pairing,

detecting any DNA:protein complexes as a measure of the presence of a base pair mismatch in said DNA duplex by utilizing an antibody specific for said base mispair recognition protein.

6. A method for detecting and localizing a base pair mismatch in a DNA duplex, comprising the steps of:

contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs;

contacting the resulting double-stranded DNA duplexes with a mispair recognition protein under conditions such that the protein forms specific complexes with mispairs,

subjecting said duplex molecules to hydrolysis with an exonuclease under conditions such that said complex blocks hydrolysis, and

determining the location of said block to hydrolysis.

7. A method for detecting and localizing a base pair mismatch in a DNA duplex, comprising the steps of

contacting at least one strand of a first DNA molecule with the complementary strand of a second DNA molecule under conditions such that base pairing occurs;

contacting the resulting double-stranded DNA duplexes with a mispair recognition protein under conditions such that the protein forms specific complexes with mispairs and thereby directs modification of at least one strand of the DNA in the resulting DNA:protein complexes in the vicinity of the DNA:protein complex, and

5,679,522

43

determining the location of the resulting DNA modification.

8. The method of claim 7, wherein said mispair recognition protein is the product of the *mutS* gene of *Escherichia coli*, an homologous protein, or a functionally equivalent protein.

9. The method of claim 7, wherein the step for modifying the DNA duplex in the vicinity of the complexed protein comprises contacting said complexes with a defined mismatch correction system or subset comprising the following purified components: *Escherichia coli* MutH, and MutL, proteins, or species variations thereof, DNA helicase II, single-stranded DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, exonuclease VII or RecJ exonuclease, ATP and one or more dideoxynucleoside triphosphates, under conditions that produce a single-stranded gap in one or both strands of the DNA duplex in the vicinity of the mismatch.

10. The method of claim 9 wherein said step for determining the location of said single-stranded gaps within said DNA duplex further includes the steps of cleaving said DNA with a single-stranded specific endonuclease, and comparing the electrophoretic mobilities of said cleaved fragments with unmodified DNA fragments.

11. The method of claim 7, wherein the step for modifying the DNA duplex in the vicinity of the complexed protein comprises:

contacting said complexes with proteins of a mismatch repair system, ATP and a divalent cation under conditions such that an endonucleolytic incision is introduced in one strand of the duplex molecule.

12. The method of claim 11, wherein the step of determining the location comprises size fractionation by electrophoretic mobility under denaturing condition relative to unmodified DNA fragments.

13. The method of claim 11, wherein said proteins of a mismatch repair system are the MutH, MutL, and MutS proteins of the *Escherichia coli* methyl-directed mismatch repair system.

14. A method for detecting and localizing a base pair mismatch within a DNA duplex, comprising the steps of:

contacting at least one strand of a first DNA molecule with a complementary strand of a second DNA molecule under conditions such that base pairing occurs;

contacting resultant duplex DNA molecules with a mispair recognition protein able to cleave DNA, under conditions such that the protein forms specific complexes with a mispair and thereby directs cleavage of at least one strand of the DNA in the resulting DNA protein complexes in the vicinity of the DNA protein complex, and

determining the location of the cleavage of the DNA.

44

15. A method for detecting and localizing A—G mispairs in a DNA duplex, comprising the steps of:

contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs;

contacting resultant duplex DNA molecules with a mispair recognition protein that recognizes A—G mispairs and an apurinic endonuclease or lyase under conditions such that in the presence of a mismatch an endonucleolytic incision is introduced in the duplex molecule, and

determining the location of the incision.

16. The method of claim 15, wherein said A—G mispair recognition protein is the product of the *mutY* gene of *Escherichia coli*.

17. The method of claim 15, wherein the analytical method comprises gel electrophoresis.

18. The method of claims 6, 7, 14, or 15, wherein the DNA molecules are obtained from the following sources: different individuals of the same species, individuals of different species, individuals of different kingdoms, different tissue types, the same tissue type in different states of growth, different cell types, cells of the same type in different states of growth, cells of the same origin in different stages of development and cells of the same type that may have undergone differential somatic mutagenesis.

19. The method of claim 18, wherein one of the DNA molecules comprises a probe sequence that has been at least partially characterized by restriction mapping or sequence analysis.

20. Assay kit for detecting and localizing individual base sequence differences within homologous regions of two DNA molecules comprising:

an aliquot of a mispair recognition protein attached to a hydroxyl radical cleaving function or attached to a DNA endonuclease.

21. The kit of claim 20, wherein the mispair recognition protein is the product of the *mutS* gene of *Escherichia coli*.

22. Assay kit for detecting and localizing an A—G mispair within a DNA duplex comprising one or more of the following components;

an aliquot of an A—G mispair recognition protein; and an aliquot of an apurinic endonuclease or lyase.

23. The kit of claim 22, wherein the A—G mispair recognition protein is the product of the *MutY* gene of *Escherichia coli*.

24. The kit of claim 20, wherein the mispair recognition protein is the product of the *mutS* gene of *Escherichia coli*.

25. The method of claim 9, wherein said exogenous dideoxynucleoside triphosphates are omitted.

* * * * *

EXHIBIT D



US005702894A

United States Patent [19]
Modrich et al.

[11] **Patent Number:** **5,702,894**
 [45] **Date of Patent:** **Dec. 30, 1997**

[54] **METHODS OF ANALYSIS AND
 MANIPULATING OF DNA UTILIZING
 MISMATCH REPAIR SYSTEMS**

[75] **Inventors:** Paul L. Modrich, Chapel Hill, N.C.;
 Shin-San Su, Newton, Mass.; Karin G.
 Au, Durham, N.C.; Robert S. Lahue,
 Northboro; Deani Lee Cooper,
 Watertown, both of Mass.; Leroy
 Worth, Jr., Durham, N.C.

[73] **Assignee:** Duke University, Durham, N.C.

[21] **Appl. No.:** 460,663

[22] **Filed:** Jun. 2, 1995

Related U.S. Application Data

[63] Continuation of Ser. No. 145,837, Nov. 1, 1993, Pat. No. 5,556,750, which is a continuation-in-part of Ser. No. 2,529, Jan. 11, 1993, abandoned, which is a continuation of Ser. No. 350,983, May 12, 1989, abandoned.

[51] **Int. Cl.⁶** C12Q 1/68; C12P 19/34;
 C07H 21/02; C07H 21/04

[52] **U.S. Cl.** 435/6; 435/91.2; 435/174;
 435/91.1; 435/7.1; 435/7.2; 435/7.9; 536/22.1;
 536/23.1; 536/24.3; 536/24.33

[58] **Field of Search** 435/6, 91.1, 91.2,
 435/174, 7.1-7.9; 536/22.1, 23.1, 24.3-24.33

[56] **References Cited**

U.S. PATENT DOCUMENTS

4,794,075	12/1988	Ford et al.	435/6
4,818,685	4/1989	Sirover et al.	435/7
5,296,231	3/1994	Yarosh et al.	424/450
5,459,039	10/1995	Modrich et al.	435/6

FOREIGN PATENT DOCUMENTS

2239456	7/1991	United Kingdom .
93022216	2/1993	WIPO .
9320233	10/1993	WIPO .
9322457	11/1993	WIPO .
9322462	11/1993	WIPO .

OTHER PUBLICATIONS

Jiricny et al. NAR 16: 7843-7853, 1988.
 Myles et al. Chemical Research in Toxicology, 2: 197-226, 1989.
 Revzin et al. Biotechniques 7: 346, 1989.
 Pang et al. J. of Bacteriol 163: 1007-1015, 1985.
 Priebe et al. J. of Bacteriol 170: 190-196, 1988.
 Chen et al. Science 237: 1197, 1987.
 Su et al. JBC 263: 6829-6835, 1988.
 Lahue et al. Science 245: 160-164, 1989.
 Lu et al. Genomics 14: 249-255, 1992.
 M. Lu et al. PNAS, 80: 4639-4634, 1983.
 Au et al. JBC 267: 12142-12148, 1992.
 N. Mack et al. J. Am Chem. Soc: 110:7572-7574, 1988.
 Lu et al. Cell 54: 805-812, 1988.
 Wu et al. J. of Bacteriology 173: 1902-1910, 1991.
 Adams et al., "The Biochemistry of the Nucleic Acids,"
 Chapman & Hall pp. 221-223 (1986).

Au et al., "Escherichia coli mutY Gene Encodes An Adenine Glycosylase Active on G-A Mispairs," *Proc. Natl. Acad. Sci. USA* 86:8877-8881 (1989).

Au et al., "Escherichia coli mutY Gene Product is Required for Specific A-G C-G Mismatch Correction," *Proc. Natl. Acad. Sci. USA* 85:9163-9166 (1988).

Au et al., "Initiation of Methyl-directed Mismatch Repair," *J. Biol. Chem.* 267:12142-12148 (1992).

Bianchi and Radding, "Insertions, Deletions and Mismatches in Heteroduplex DNA Made by RecA Protein," *Cell* 35:511-520 (1983).

Chen and Sigman, "Chemical Conversion of a DNA-Binding Protein Into a Site-Specific Nuclease," *Science* 237:1197-1201 (1987).

Cooper et al., "Methyl-Directed Mismatch Repair is Bidirectional," *J. Biol. Chem.* 268:11823-11829 (1993).

Cotton et al., "Reactivity of Cytosine and Thymine In Single-Base-Pair Mismatches with Hydroxylamine and Osmium Tetroxide and Its Application to the Study of Mutations," *Proc. Natl. Acad. Sci. USA* 85:4397-4401 (1988).

Cox and Lehman, "recA protein of *Escherichia coli* promotes branch migration, a kinetically distinct phase of DNA strand exchange," *Proc. Natl. Acad. Sci. USA* 78:3433-3437 (1981).

Dasgupta and Radding, "Polar Branch Migration Promoted by recA Protein: Effect of Mismatched Base Pairs," *Proc. Natl. Acad. Sci. USA* 79:762-766 (1982).

Fang and Modrich, "Human Strand-Specific Mismatch Repair Occurs by a Bidirectional Mechanism Similar to That of the Bacterial Reaction," *J. Biol. Chem.* 268:11838-11844 (1993).

Feinstein and Low, "Hyper-Recombining Recipient Strains in Bacterial Conjugation," *Genetics* 113:13-33 (1986).

Fujii and Shimada, "Isolation and Characterization of cDNA Clones Derived from the Divergently Transcribed Gene in the Region Upstream from the Human Dihydrofolate Reductase Gene," *J. Biol. Chem.* 264:10057-10064 (1989).

Grilley et al., "Isolation and Characterization of the *Escherichia coli* mutL Gene Product," *J. Biol. Chem.* 264:1000-1004 (1989).

(List continued on next page.)

Primary Examiner—W. Gary Jones

Assistant Examiner—Dianne Rees

Attorney, Agent, or Firm—Lyon & Lyon LLP

[57] **ABSTRACT**

A diagnostic method for detecting a base pair mismatch in a DNA duplex, comprising the steps of contacting at least one strand of a first DNA molecule with the complementary strand of a second DNA molecule under conditions such that base pairing occurs contacting a DNA duplex potentially containing a base pair mismatch with a mismatch recognition protein under conditions suitable for the protein to form a specific complex only with the DNA duplex having a base pair mismatch, and not with a DNA duplex lacking a base pair mismatch, and detecting any complex as a measure of the presence of a base pair mismatch in the DNA duplex.

8 Claims, 8 Drawing Sheets

5,702,894

Page 2

OTHER PUBLICATIONS

- Grilley et al., "Mechanisms of DNA-Mismatch Correction," *Mutation Research* 236:253-267 (1990).
- Grilley et al., "Bidirectional Excision in Methyl-Directed Mismatch Repair," *J. Biol. Chem.* 268:11830-11837 (1993).
- Hennighausen and Lubon, "Interaction of Protein With DNA In Vitro," *Guide to Molecular Cloning Techniques*, Berger and Kimmel eds., 152:721-735 (1987).
- Holmes et al., "Strand-specific Mismatch Correction In Nuclear Extracts of Human and Drosophila Melanogaster Cell Lines," *Proc. Natl. Acad. Sci. USA* 87:5837-5841 (1990).
- Jiricny et al., "Mismatch-containing Oligonucleotide Duplexes Bound By the *E. coli* mutS-encoded Protein," *Nucleic Acids Research* 16:7843-7853 (1988).
- Lahue et al., "Requirement for d(GATC) Sequences in *Escherichia coli* mutHLS Mismatch Correction," *Proc. Natl. Acad. Sci. USA* 84:1482-1486 (1987).
- Lahue and Modrich, "DNA Mismatch Correction in a Defined System," *Science* 245:160-164 (1989).
- Lahue and Modrich, "Methyl-directed DNA Mismatch Repair in *Escherichia coli*," *Mutation Research* 198:37-43 (1988).
- Lu et al., "Repair of DNA Base-pair Mismatches in Extracts of *Escherichia coli*," *Cold Spring Harbor Laboratory, Cold Spring Harbor Symposia on Quantitative Biology*, XLIX:589-596 (1984).
- Lu and Chang, "Repair of Single Base-Pair Transversion Mismatches of *Escherichia coli* in Vitro: Correction of Certain A/G Mismatches is Independent of dam Methylation and Host mutHLS Gene Functions," *Genetics* 118:593-600 (1988).
- Lu and Chang, "A Novel Nucleotide Excision Repair for The Conversion of An A/G Mismatch to C/G Base Pair in *E. coli*," *Cell* 54:805-812 (1988).
- Lu, "Influence of GATC Sequences on *Escherichia coli* DNA Mismatch Repair In Vitro," *J. Bacteriology* 169:1254-1259 (1987).
- Lu and Hsu, "Detection of Single DNA Base Mutations with Mismatch Repair Enzymes," *Genomics* 14:249-255 (1992).
- Lu et al., "Methyl-directed Repair of DNA Base-pair Mismatches in Vitro," *Proc. Natl. Acad. Sci. USA* 80:4639-4643 (1983).
- Marx, "DNA Repair Comes Into Its Own," *Science* 266:728-730 (1994).
- Modrich, *Molecular Mechanisms of DNA-Protein Interaction*, 1986, NIH Grant, Abstract (Source: CRISP) (vol. #, p. # not applicable).
- Modrich, "Mechanisms and Biological Effects on Mismatch Repair," *Ann. Rev. Genet.* 25:229-253 (1991).
- Modrich et al., "DNA Mismatch Correction," *Ann. Rev. Biochem.* 56:435-466 (1987).
- Modrich, "Mismatch, Repair, Genetic Stability and Cancer," *Science* 266:1959-1960 (1994).
- Modrich, "Methyl-directed DNA Mismatch Correction," *J. Biol. Chem.* 264:6597-6600 (1989).
- Myers et al., "Detection of Single Based Substitutions by Ribonuclease Cleavage at Mismatches in RNA:DNA Duplexes," *Science* 230:1242-1246 (1985).
- Nelson et al., "Genomic Mismatch Scanning A New Approach to Genetic Linkage Mapping," *Nature Genetics* 4:11-19 (1993).
- Pang et al., "Identification and Characterization of the mutL and mutS Gene Products of *Salmonella typhimurium* LT2," *J. Bacteriology* 163:1007-1015 (1985).
- Petit et al., "Control of Large Chromosomal Duplications in *Escherichia coli* by the Mismatch Repair System," *Genetics* 129:327-332 (1991).
- Priebe et al., "Nucleotide Sequence of the hexA Gene for DNA Mismatch Repair in *Streptococcus pneumoniae* and Homology of hexA to mutS of *Escherichia coli* and *Salmonella typhimurium*," *J. Bacteriology* 170:190-196 (1988).
- Quinones et al., "Expression of the *Escherichia coli* dna Q (mutD) Gene is Inducible," *Mol. Gene Genet.* 211:106-112 (1988).
- Rayssiguier et al., "The Barrier to Recombination Between *Escherichia coli* and *Salmonella typhimurium* is Disrupted in Mismatch-Repair Mutants," *Nature* 342:396-401 (1989).
- Reenan and Kolodner, "Isolation and Characterization of Two *Saccharomyces cerevisiae* Genes Encoding Homologs of the Bacterial HexA and MutS Mismatch Repair Proteins," *Genetics* 132:963-973 (1992).
- Shen and Huang, "Effect of Base Pair Mismatches on Recombination Via the RecBCD Pathway," *Mol. Gen. Genet.* 218:358-360 (1989).
- Su and Modrich, "*Escherichia coli* mutS-encoded Protein binds to Mismatched DNA Base Pairs," *Proc. Natl. Acad. Sci. USA* 83:5057-5061 (1986).
- Su et al., "Gap Formation is Associated With Methyl-Directed Mismatch Correction Under Conditions of Restricted DNA Synthesis," *Genome* 31:104-111 (1989).
- Su et al., "Mispair Specificity of Methyl-directed DNA Mismatch Correction in Vitro," *J. Biol. Chem.* 263:6829-6835 (1988).
- Wilchle et al., *Analytical Biochem.* 171:1-32 (1988).
- Welsh et al., "Isolation and Characterization of the *Escherichia coli* mutH Gene Product," *J. Biol. Chem.* 262:15624-15629 (1987).

U.S. Patent

Dec. 30, 1997

Sheet 1 of 8

5,702,894

V 5'-AAGCTTTCGAG Hind III
C 3'-TTCGAGAGCTC Xho I

FIG. 1.

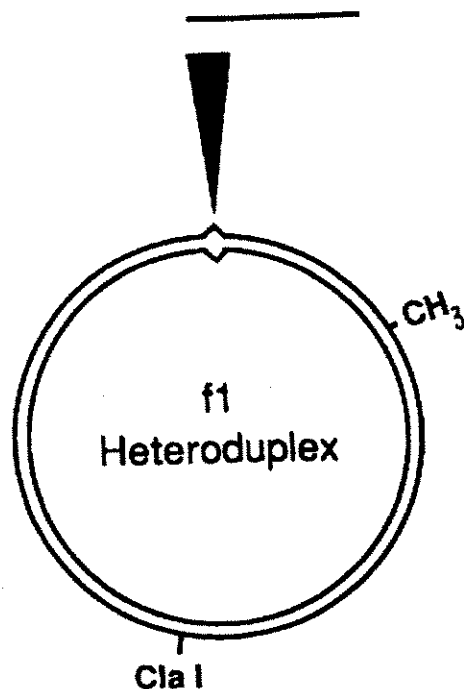
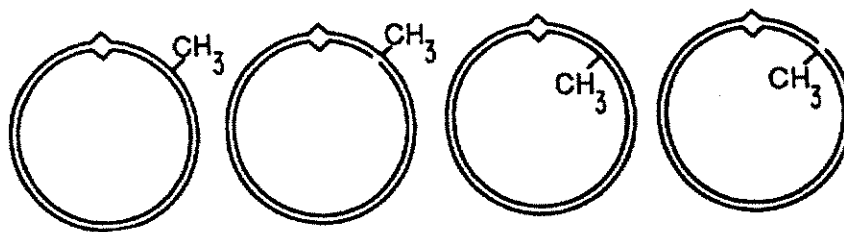


FIG. 4.



Reaction conditions	Repair (fmol/20 min)			
Complete	15 (<1)	17 (<1)	8 (<1)	10 (<1)
- Mut H	<1	18	1	9
- Mut L	<1	<1	<1	<1
- Mut S	<1	<1	<1	1
- SSB	2	<1	<1	<1
- pol III holoenzyme	<1	<1	<1	<1

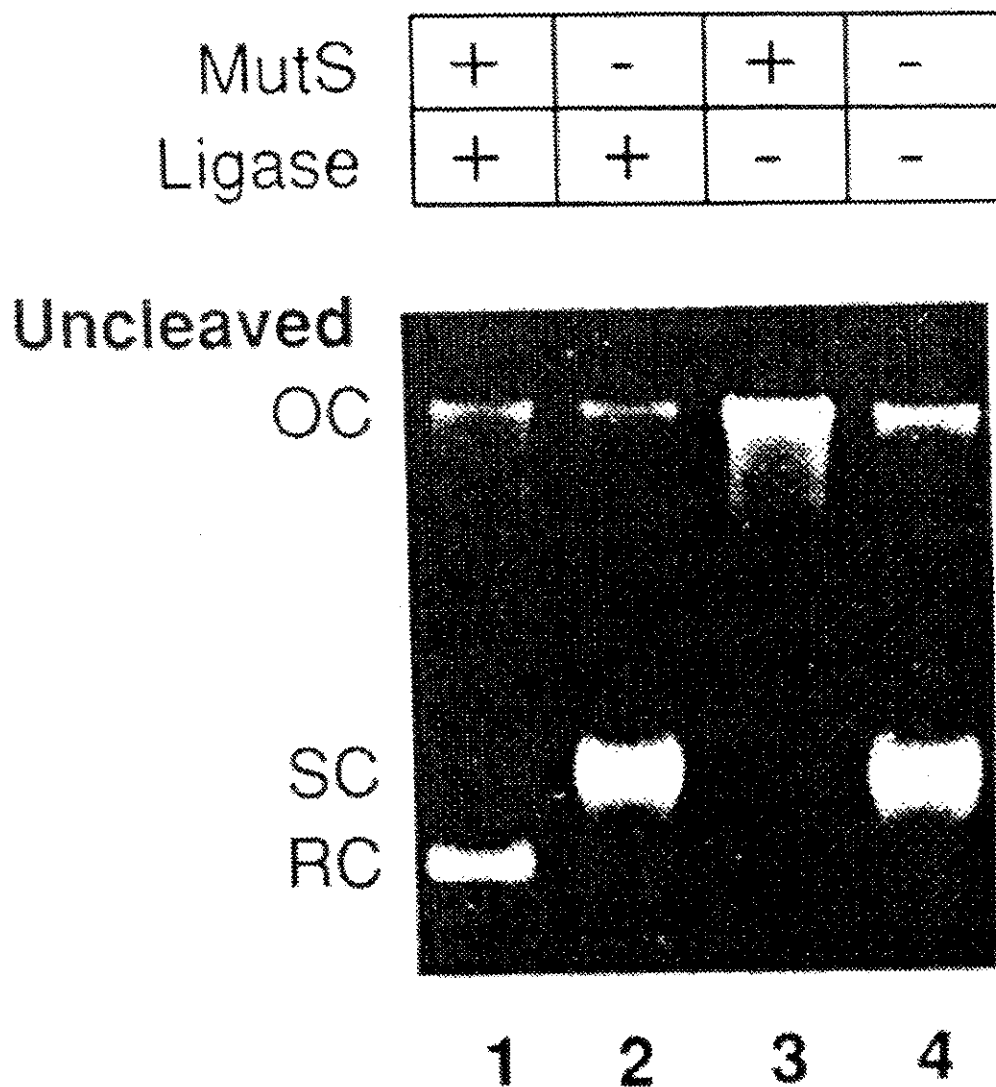
U.S. Patent

Dec. 30, 1997

Sheet 2 of 8

5,702,894

FIG. 2A.



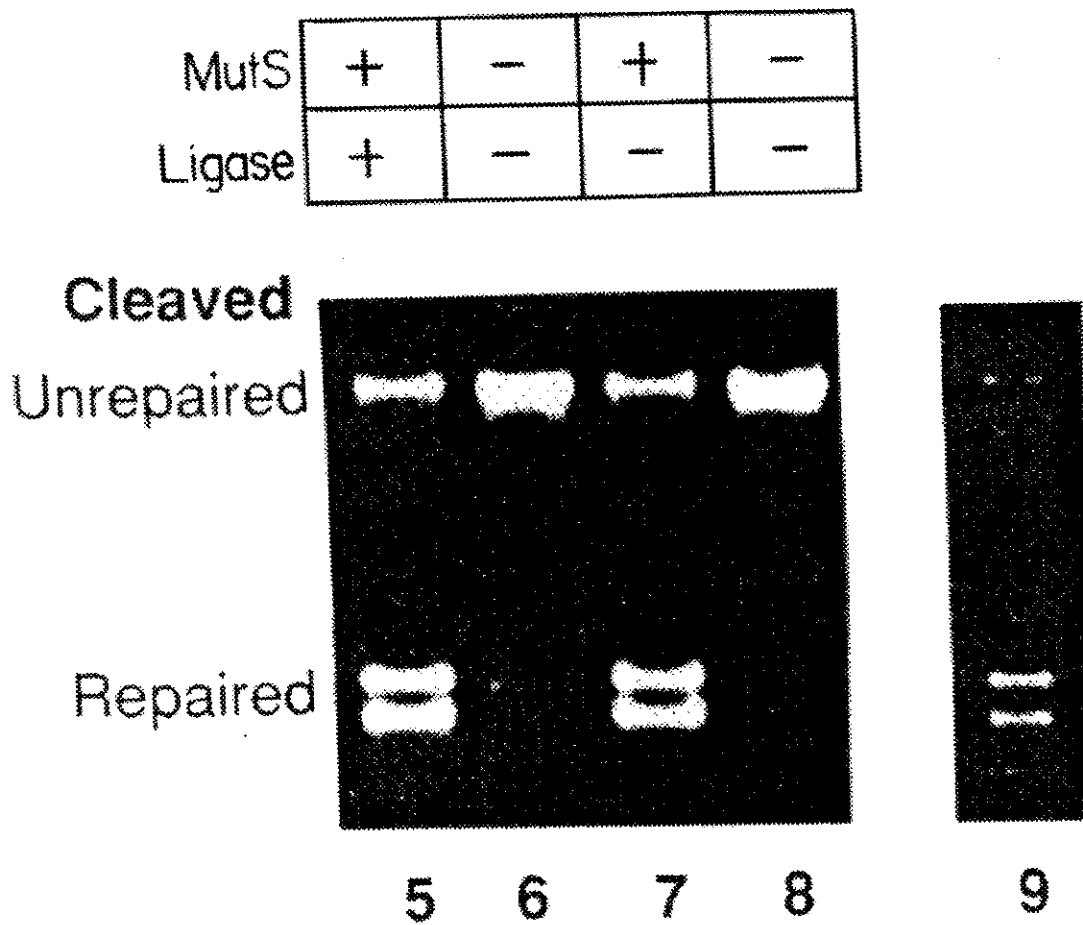
U.S. Patent

Dec. 30, 1997

Sheet 3 of 8

5,702,894

FIG. 2B.



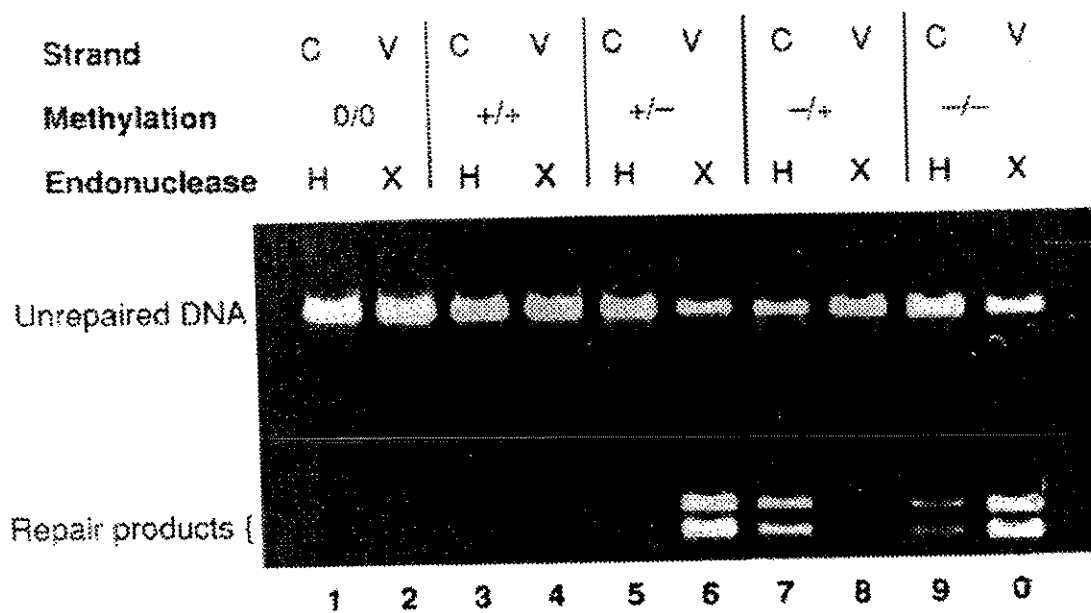
U.S. Patent

Dec. 30, 1997

Sheet 4 of 8

5,702,894

FIG. 3.

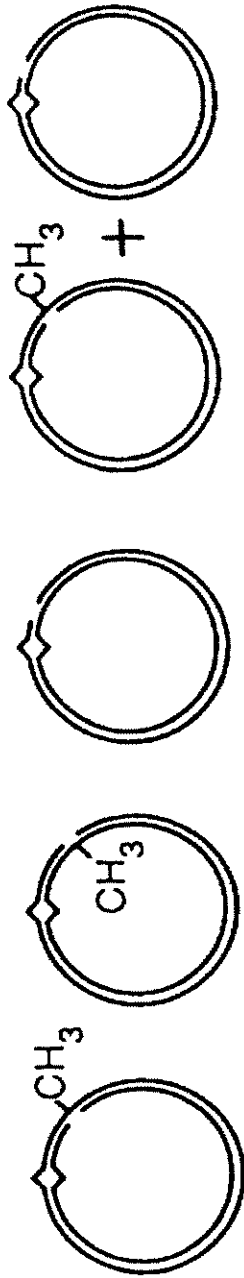


U.S. Patent

Dec. 30, 1997

Sheet 5 of 8

5,702,894



Repair (fmol/20 min)

Ligase Muth

		19 (<1)	9 (<1)	11 (<1)	19 (<1)	9 (<1)
—	—	19 (<1)	9 (<1)	11 (<1)	19 (<1)	9 (<1)
+	—	2	<1	1	2	1
+	+	20	7	2	15	1

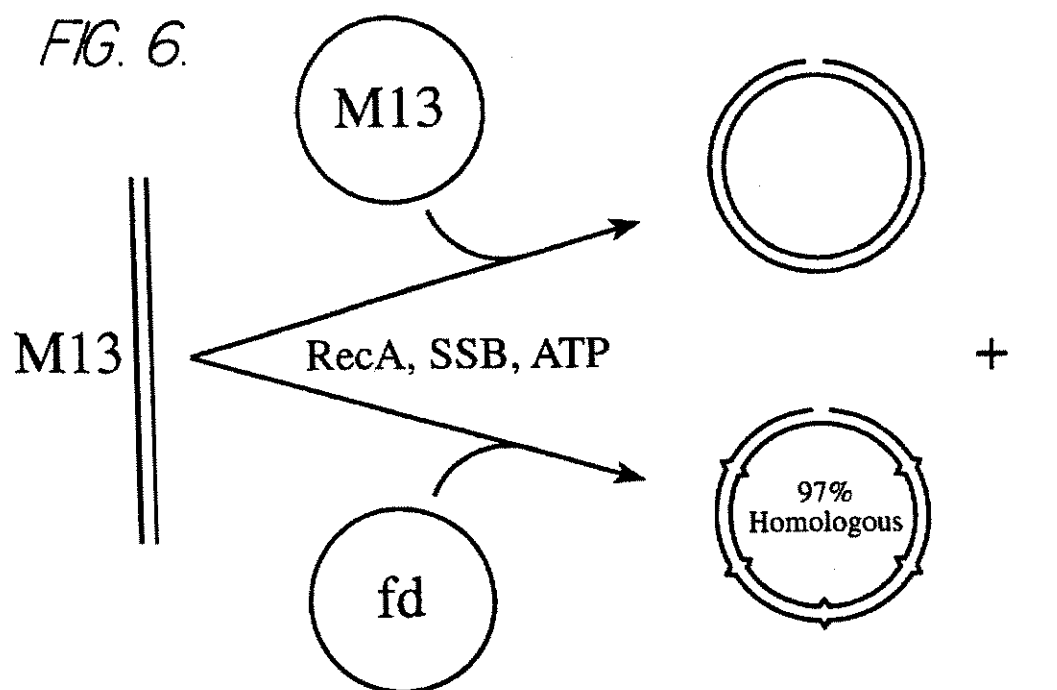
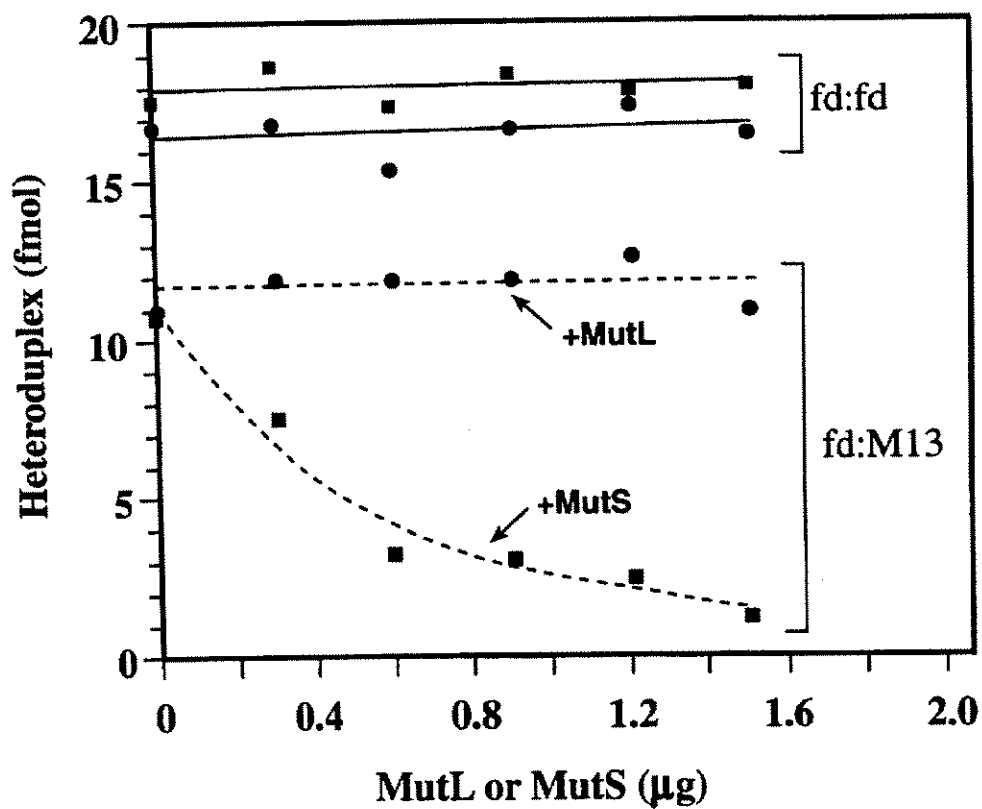
FIG. 5.

U.S. Patent

Dec. 30, 1997

Sheet 6 of 8

5,702,894

*FIG. 7.*

U.S. Patent

Dec. 30, 1997

Sheet 7 of 8

5,702,894

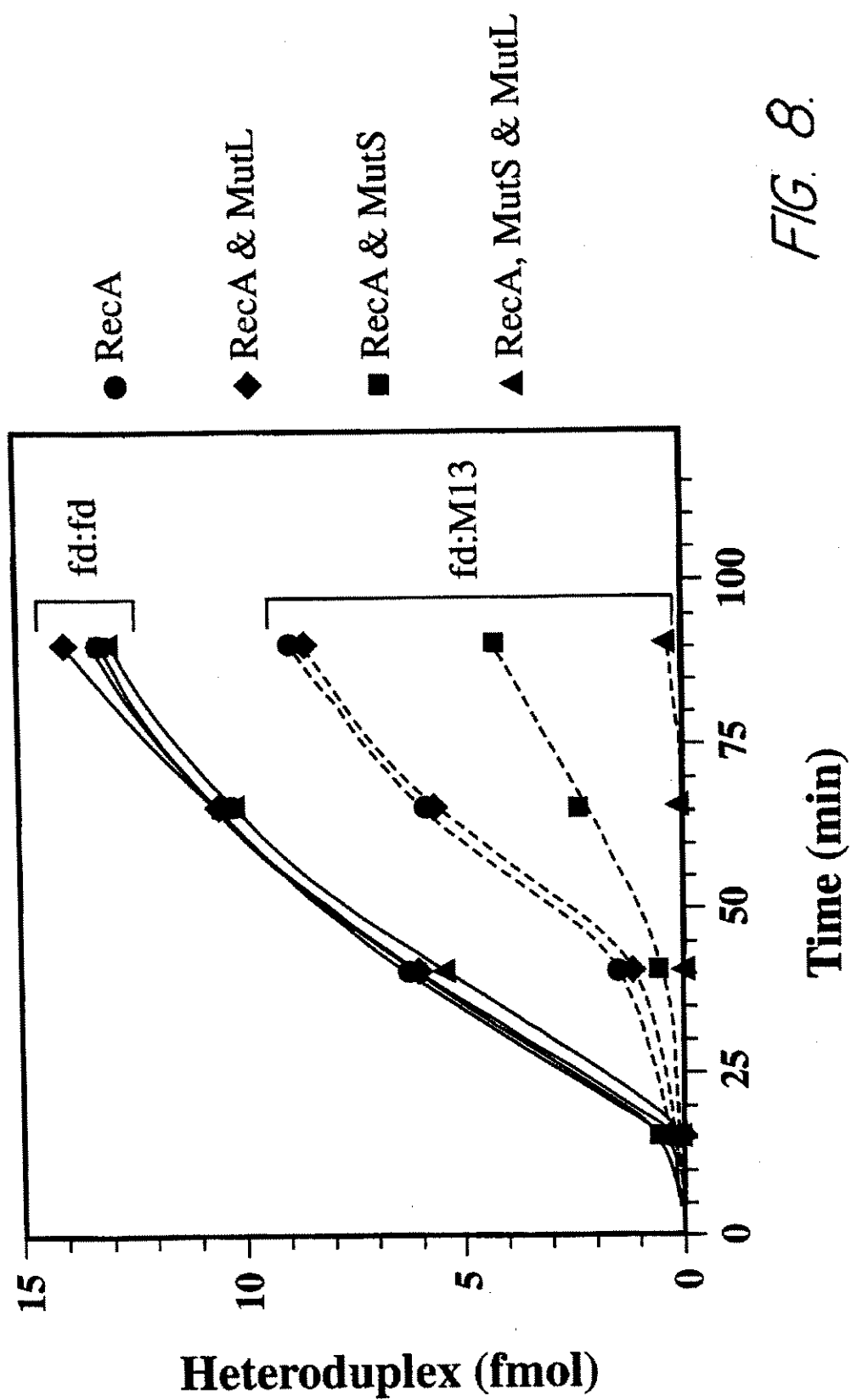


FIG. 8.

U.S. Patent

Dec. 30, 1997

Sheet 8 of 8

5,702,894

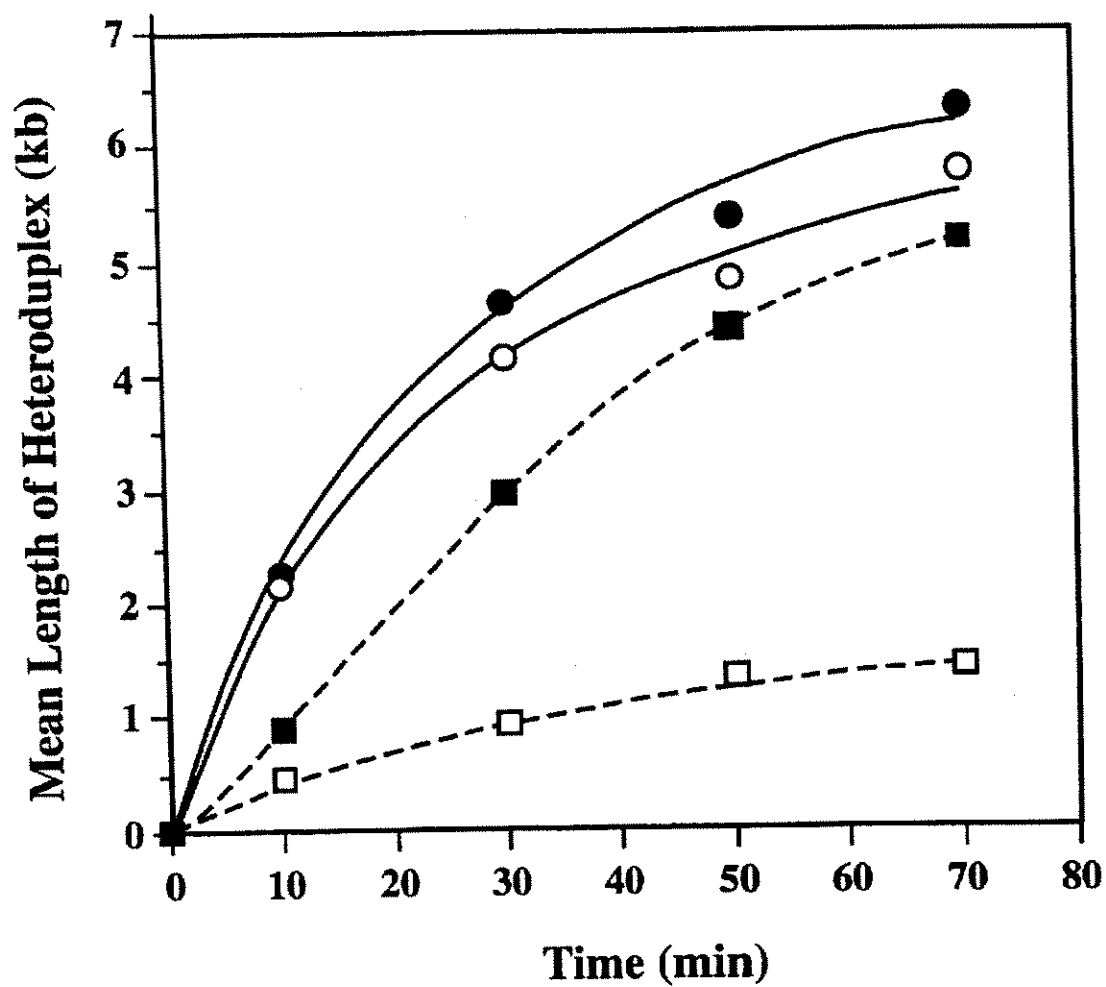


FIG. 9.

5,702,894

1

METHODS OF ANALYSIS AND MANIPULATING OF DNA UTILIZING MISMATCH REPAIR SYSTEMS

BACKGROUND OF THE INVENTION

This application is a continuation of Modrich et. al. U.S. Ser. No. 08/145,837 filed Nov. 1, 1993, now U.S. Pat. No. 5,556,750 which is a continuation-in-part of Modrich et al., U.S. Ser. No. 08/002,529, filed Jan. 11, 1993, now abandoned entitled "Methods For Mapping Genetic Mutations" which is a continuation of U.S. Ser. No. 07/350,983, filed May 12, 1989, now abandoned entitled, "Methods For Mapping Genetic Mutations", both hereby incorporated by reference herein, including drawings.

DESCRIPTION

This work was supported by the U.S. government, namely Grant No. GM23719. The U.S. government may have rights in this invention.

FIELD OF THE INVENTION

The present invention relates to methods for mapping genetic differences among deoxyribonucleic acid ("DNA") molecules, especially mutations involving a difference in a single base between the base sequences of two homologous DNA molecules.

The following is a discussion of relevant art, none of which is admitted to be prior art to the appended claims.

Mapping of genetic differences between individuals is of growing importance for both forensic and medical applications. For example, DNA "fingerprinting" methods are being applied for identification of perpetrators of crimes where even small amounts of blood or sperm are available for analysis. Biological parents can also be identified by comparing DNAs of a child and a suspected parent using such means. Further, a number of inherited pathological conditions may be diagnosed before onset of symptoms, even in utero, using methods for structural analyses of DNA. Finally, it is notable that a major international effort to physically map and, ultimately, to determine the sequence of bases in the DNA encoding the entire human genome is now underway and gaining momentum in both institutional and commercial settings.

DNA molecules are linear polymers of subunits called nucleotides. Each nucleotide comprises a common cyclic sugar molecule, which in DNA is linked by phosphate groups on opposite sides to the sugars of adjoining nucleotides, and one of several cyclic substituents called bases. The four bases commonly found in DNAs from natural sources are adenine, guanine, cytosine and thymine, hereinafter referred to as A, G, C and T, respectively. The linear sequence of these bases in the DNA of an individual encodes the genetic information that determines the heritable characteristics of that individual.

In double-stranded DNA, such as occurs in the chromosomes of all cellular organisms, the two DNA strands are entwined in a precise helical configuration with the bases projecting inward and so aligned as to allow interactions between bases from opposing strands. The two strands are held together in precise alignment mainly by hydrogen bonds which are permitted between bases by a complementarity of structures of specific pairs of bases. This structural complementarity is determined by the chemical natures and locations of substituents on each of the bases. Thus, in double-stranded DNA, normally each A on one strand pairs

2

with a T from the opposing strand, and, likewise, each G with an opposing C.

When a cell undergoes reproduction, its DNA molecules are replicated and precise copies are passed on to its descendants. The linear base sequence of a DNA molecule is maintained in the progeny during replication in the first instance by the complementary base pairings which allow each strand of the DNA duplex to serve as a template to align free nucleotides with its polymerized nucleotides. The complementary nucleotides so aligned are biochemically polymerized into a new DNA strand with a base sequence that is entirely complementary to that of the template strand.

Occasionally, an incorrect base pairing does occur during replication, which, after further replication of the new strand, results in a double-stranded DNA offspring with a sequence containing a heritable single base difference from that of the parent DNA molecule. Such heritable changes are called genetic mutations, or more particularly in the present case, "single base pair" or "point" mutations. The consequences of a point mutation may range from negligible to lethal, depending on the location and effect of the sequence change in relation to the genetic information encoded by the DNA.

The bases A and G are of a class of compounds called purines, while T and C are pyrimidines. Whereas the normal base pairings in DNA (A with T, G with C) involve one purine and one pyrimidine, the most common single base mutations involve substitution of one purine or pyrimidine for the other (e.g., A for G or C for T or vice versa), a type of mutation referred to as a "transition". Mutations in which a purine is substituted for a pyrimidine, or vice versa, are less frequently occurring and are called "transversions". Still less common are point mutations comprising the addition or loss of a small number (1, 2 or 3) of nucleotides arising in one strand of a DNA duplex at some stage of the replication process. Such mutations are called small "insertions" or "deletions", respectively, and are also known as "frameshift" mutations in the case of insertion/deletion of one of two nucleotides, due to their effects on translation of the genetic code into proteins. Mutations involving larger sequence rearrangement also do occur and can be important in medical genetics, but their occurrences are relatively rare compared to the classes summarized above.

Mapping of genetic mutations involves both the detection of sequence differences between DNA molecules comprising substantially identical (i.e., homologous) base sequences, and also the physical localization of those differences within some subset of the sequences in the molecules being compared. In principle, it is possible to both detect and localize limited genetic differences, including point mutations within genetic sequences of two individuals, by directly comparing the sequences of the bases in their DNA molecules.

Other methods for detecting differences between DNA sequences have been developed. For example, some pairs of single-stranded DNA fragments with sequences differing in a single base may be distinguished by their different migration rates in electric fields, as in denaturing gradient gel electrophoresis.

DNA restriction systems found in bacteria for example, comprise proteins which generally recognize specific sequences in double-stranded DNA composed of 4 to 6 or more base pairs. In the absence of certain modifications (e.g., a covalently attached methyl group) at definite positions within the restriction recognition sequence, endonuclease components of the restriction system will cleave both

5,702,894

3

strands of a DNA molecule at specific sites within or near the recognition sequence. Such short recognition sequences occur by chance in all natural DNA sequences, once in every few hundred or thousand base pairs, depending on the recognition sequence length. Thus, digestion of a DNA molecule with various restriction endonucleases, followed by analyses of the sizes of the resulting fragments (e.g., by gel electrophoresis), may be used to generate a physical map ("fingerprint") of the locations in a DNA molecule of selected short sequences.

Comparisons of such restriction maps of two homologous DNA sequences can reveal differences within those specific sequences that are recognized by those restriction enzymes used in the available maps. Restriction map comparisons may localize any detectable differences within limits defined ultimately by the resolving power of DNA fragment size determination, essentially within about the length of the restriction recognition sequence under certain conditions of gel electrophoresis.

In practice, selected heritable differences in restriction fragment lengths (i.e., restriction fragment length polymorphisms, "RFLP"s) have been extremely useful, for instance, for generating physical maps of the human genome on which genetic defects may be located with a relatively low precision of hundreds or, sometimes, tens of thousands of base pairs. Typically, RFLPs are detected in human DNA isolated from small tissue or blood samples by using radioactively labeled DNA fragments complementary to the genes of interest. These "probes" are allowed to form DNA duplexes with restriction fragments of the human DNA after separation by electrophoresis, and the resulting radioactive duplex fragments are visualized by exposure to photographic (e.g., X-ray sensitive) film, thereby allowing selective detection of only the relevant gene sequences amid the myriad of others in the genomic DNA.

When the search for DNA sequence differences can be confined to specific regions of known sequence, the recently developed "polymerase chain reaction" ("PCR") technology can be used. Briefly, this method utilizes short DNA fragments complementary to sequences on either side of the location to be analyzed to serve as points of initiation for DNA synthesis (i.e., "primers") by purified DNA polymerase. The resulting cyclic process of DNA synthesis results in massive biochemical amplification of the sequences selected for analysis, which then may be easily detected and, if desired, further analyzed, for example, by restriction mapping or direct DNA sequencing methods. In this way, selected regions of a human gene comprising a few kbp may be amplified and examined for sequence variations.

Another known method for detecting and localizing single base differences within homologous DNA molecules involves the use of a radiolabeled RNA fragment with base sequence complementary to one of the DNAs and a nuclease that recognizes and cleaves single-stranded RNA. The structure of RNA is highly similar to DNA, except for a different sugar and the presence of uracil (U) in place of T; hence, RNA and DNA strands with complementary sequences can form helical duplexes ("DNA:RNA hybrids") similar to double-stranded DNA, with base pairing between A's and U's instead of A's and T's. It is known that the enzyme ribonuclease A ("RNase A") can recognize some single pairs of mismatched bases (i.e., "base mismatches") in DNA:RNA hybrids and can cleave the RNA strand at the mismatch site. Analysis of the sizes of the products resulting from RNase A digestion allows localization of single base mismatches, potentially to the precise sequence position, within lengths of homologous sequences determined by the limits of reso-

4

lution of the RNA sizing analysis (Myers, R. M. et al., 1985, Science, 230, 1242-1246). RNA sizing is performed in this method by standard gel electrophoresis procedures used in DNA sequencing.

S1 nuclease, an endonuclease specific for single-stranded nucleic acids, can recognize and cleave limited regions of mismatched base pairs in DNA:DNA or DNA:RNA duplexes. A mismatch of at least about 4 consecutive base pairs actually is generally required for recognition and cleavage of a duplex by S1 nuclease.

Ford et al., (U.S. Pat. No. 4,794,075) disclose a chemical modification procedure to detect and localize mispaired guanines and thymidines and to fractionate a pool of hybrid DNA from two samples obtained from related individuals. Carbodiimide is used to specifically derivatize unpaired G's and T's, which remain covalently associated with the DNA helix.

The present invention concerns use of proteins that function biologically to recognize mismatched base pairs in double-stranded DNA (and, therefore, are called "mismatch recognition proteins") and their application in defined systems for detecting and mapping point mutations in DNAs. Accordingly, it is an object of the present invention to provide methods for using such mismatch recognition proteins, alone or in combination with other proteins, for detecting and localizing base pair mismatches in duplex DNA molecules, particularly those DNAs comprising several kbp, and manipulating molecules containing such mismatches. Additionally, it is an object of this invention to develop modified forms of mismatch recognition proteins to further simplify methods for identifying specific bases which differ between DNAs. The following is a brief outline of the art regarding mismatch recognition proteins and systems, none of which is admitted to be prior art to the present invention.

Enzymatic systems capable of recognition and correction of base pairing errors within the DNA helix have been demonstrated in bacteria, fungi and mammalian cells, but the mechanisms and functions of mismatch correction are best understood in *Escherichia coli*. One of the several mismatch repair systems that have been identified in *E. coli* is the methyl-directed pathway for repair of DNA biosynthetic errors. The fidelity of DNA replication in *E. coli* is enhanced 100-1000 fold by this post-replication mismatch correction system. This system processes base pairing errors within the helix in a strand-specific manner by exploiting patterns of DNA methylation. Since DNA methylation is a post-synthetic modification, newly synthesized strands temporarily exist in an unmethylated state, with the transient absence of adenine methylation on GATC sequences directing mismatch correction to new DNA strands within the hemimethylated duplexes.

In vivo analyses in *E. coli* have shown that selected examples of each of the different mismatches are subject to correction with different efficiencies. G-T, A-C, G-G and A-A mismatches are typically subject to efficient repair. A-G, C-T, T-T and C-C are weaker substrates, but well repaired exceptions exist within this class. The sequence environment of a mismatched base pair may be an important factor in determining the efficiency of repair in vivo. The mismatch correction system is also capable in vivo of correcting differences between duplexed strands involving a single base insertion or deletion. Further, genetic analyses have demonstrated that the mismatch correction process requires intact genes for several proteins, including the products of the mutH, mutL, and mutS genes, as well as DNA

5,702,894

5

helicase II and single-stranded DNA binding protein (SSB). The following are further examples of art discussing this subject matter.

Lu et al., 80 *Proc. Natl. Acad. Sci. USA* 4639, 1983 disclose the use of a soluble *E. coli* system to support mismatch correction in vitro.

Pans et al., 163 *J. Bact.* 1007, 1985 disclose cloning of the mutS and mutL genes of *Salmonella typhimurium*.

The specific components of the *E. coli* mismatch correction system have been isolated and the biochemical functions determined. Preparation of MutS protein substantially free of other proteins has been reported (Su and Modrich, 1986, *Proc. Nat. Acad. Sci. USA*, 84, 5057-5061, which is hereby incorporated herein by reference). The isolated MutS protein was shown to recognize four of the eight possible mismatched base pairs (specifically, G-T, A-C, A-G and C-T mispairs).

Su et al., 263 *J. Biol. Chem.* 6829, 1988 disclose that the mutS gene product binds to each of the eight base pair mismatches and does so with differential efficiency.

Jiricny et al., 16 *Nucleic Acids Research* 7843, 1988 disclose binding of the mutS gene product of *E. coli* to synthetic DNA duplexes containing mismatches to correlate recognition of mispairs and efficiency of correction in vivo. Nitrocellulose filter binding assays and band-shift assays were utilized.

Welsh et al., 262 *J. Biol. Chem.* 15624, 1987 purified the product of the MutH gene to near homogeneity and demonstrated the MutH gene product to be responsible for d(GATC) site recognition and to possess a latent endonuclease that incises the unmethylated strand of hemimethylated DNA 5' to the G of d(GATC) sequences.

Au et al., 267 *J. Biol. Chem.* 12142, 1992 indicate that activation of the MutH endonuclease requires MutS, MutL, and ATP.

Grilley et al. 264 *J. Biol. Chem.* 1000, 1989 purified the *E. coli* mutL gene product to near homogeneity and indicate that the mutL gene product interacts with MutS heteroduplex DNA complex.

Lahue et al., 245 *science* 160, 1989 delineate the components of the *E. coli* methyl-directed mismatch repair system that function in vitro to correct seven of the eight possible base pair mismatches. Such a reconstituted system consists of MutH, MutL, and MutS proteins, DNA helicase II, single-strand DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, DNA ligase, ATP, and the four deoxyribonucleoside triphosphates.

Suet et al., 31 *Genome* 104, 1989 indicate that under conditions of restricted DNA synthesis, or limiting concentration of dNTPs, or by supplementing a reaction with a ddNTP, there is the formation of excision tracts consisting of single-stranded gaps in the region of the molecule containing a mismatch and a d(GATC) site.

Grilley et al. 268 *J. Biol. Chem.* 11830, 1993, indicate that excision tracts span the shorter distance between a mismatch and the d(GATC) site, indicating a bidirectional capacity of the methyl-directed system.

Holmes et al., 87 *Proc. Natl. Acad. Sci. USA*, 5837, 1990, disclose nuclear extracts derived from *Hela* and *Drosophila melanogaster* K_c cell lines to support strand mismatch correction in vitro.

Cooper et al., 268 *J. Biol. Chem.*, 11823, 1993, describe a role for RecJ and Exonuclease VII as a 5' to 3' exonuclease in a mismatch repair reaction. In reconstituted systems such as a 5' to 3' exonuclease function had been provided by certain preparations of DNA polymerase III holoenzyme.

6

Au et al., 86 *Proc. Natl. Acad. Sci. USA* 8877, 1989 describe purification of the MutY gene product of *E. coli* to near homogeneity, and state that the MutY protein is a DNA glycosylase that hydrolyzes the glycosyl bond linking a mispaired adenine (G-A) to deoxyribose. The MutY protein, an apurinic endonuclease, DNA polymerase I, and DNA ligase were shown to reconstitute G-A to G-C mismatch correction in vitro.

A role for the *E. coli* mismatch repair system in controlling recombination between related but non allelic sequences has been indicated (Feinstein and Low, 113 *Genetics* 13, 1986; Rayssiguier, 342 *Nature* 396, 1989; Shen, 218 *Mol. Gen. Genetics* 358, 1989; Petit, 129 *Genetics* 327, 1991). The frequency of crossovers between sequences which differ by a few percent or more at the base pair level are rare. In bacterial mutants deficient in methyl-directed mismatch repair, the frequency of such events increases dramatically. The largest increases are observed in MutS and MutL deficient strains. (Rayssiguier, supra; and Petit, supra.)

Nelson et al., 4 *Nature Genetics* 11, 1993, disclose a genomic mismatch (GMS) method for genetic linkage analysis. The method allows DNA fragments from regions of identity-by-descent between two relatives to be isolated based on their ability to form mismatch-free hybrid molecules.

The method consists of digesting DNA from the two sources with a restriction endonuclease that produces protruding 3' ends. The protruding 3' ends provide some protection from exonuclease III, which is used in later steps. The two sources are distinguished by methylating the DNA from only one source. Molecules from both sources are denatured and reannealed, resulting in the formation of four types of duplex molecules: homohybrids formed from strands derived from the same source and heterohybrids consisting of DNA strands from different sources. Heterohybrids can either be mismatch free or contain base-pair mismatches, depending on the extent of identity of homologous regions.

Homohybrids are distinguished from heterohybrids by use of restriction endonucleases that cleave at fully methylated or unmethylated GATC sites. Homohybrids are cleaved to smaller duplex molecules, while heterohybrid are resistant to cleavage. Heterohybrids containing a mismatch (es) are distinguished from mismatch free molecules by use of the *E. coli* methyl-directed mismatch repair system. The combination of three proteins of the methyl-directed mismatch repair system MutH, MutL, and MutS along with ATP introduce a single-strand nick on the unmethylated strand at GATC sites in duplexes that contain a mismatch. Heterohybrids that do not contain a mismatch are not nicked. All molecules are then subject to digestion by Exonuclease III (Exo III), which can initiate digestion at a nick, a blunt end or a 5' overhang, to produce single-stranded gaps. Only mismatch free heterohybrids are not subject to attack by Exo III, all other molecules have single-stranded gaps introduced by the enzyme. Molecules with single-stranded regions are removed by absorption to benzoylated naphthoylated DEAE cellulose. The remaining molecules consist of mismatch-free heterohybrids which may represent regions of identity by descent.

SUMMARY OF THE INVENTION

Applicant has determined that a single DNA base mispair recognition protein can form specific complexes with any of the eight possible mismatched base pairs embedded in an otherwise homologous DNA duplex. It has also been

5,702,894

7

revealed that another mispair recognition protein can recognize primarily one specific base pair mismatch, A-G, and in so doing, it chemically modifies a nucleotide at the site of the mispair. In addition, defined *in vitro* systems have been established for carrying out methyl-directed mismatch repair processes. Accordingly, the present invention features the use of such mispair recognition proteins and related correction system components to detect and to localize point mutations in DNAs. In addition the invention concerns methods for the analysis and manipulation of populations of DNA duplex molecules potentially containing base pair mismatches through the use of all or part of defined mismatch repair systems.

The invention utilizes five basic methods for heteroduplex mapping analysis, and manipulation: (i) binding of a mismatch recognition protein, e.g., MutS to DNA molecules containing one or more mispairs; (ii) cleavage of a heteroduplex in the vicinity of a mismatch by a modified form of a mismatch recognition protein; (iii) mismatch-provoked cleavage at one or more GATC sites via a mismatch repair system dependent reaction, e.g., MthLS; (iv) formation of a mismatch-provoked gap in heteroduplex DNA via reactions of a mismatch repair system and (v) labelling of mismatch-containing nucleotides with a nucleotide analog, e.g., a biotinylated nucleotide, using a complete mismatch repair system.

For clarity in the following discussion, it should be noted that certain distinctions exist related to the fact that some proteins that recognize DNA base mispairs are merely DNA binding proteins, while others modify the DNA as a consequence of mispair recognition. Notwithstanding the fact that in the latter situation the protein modifying the DNA may be associated with the DNA only transiently, hereinafter, whether a mispair recognition protein is capable of DNA binding only or also of modifying DNA, whenever it is said that a protein recognizes a DNA mispair, this is equivalent to saying that it "forms specific complexes with" or "binds specifically to" that DNA mispair in double-stranded DNA. In the absence of express reference to modification of DNA, reference to DNA mispair recognition does not imply consequent modification of the DNA. Further, the phrase "directs modification of DNA" includes both cases wherein a DNA mispair recognition protein has an inherent DNA modification function (e.g., a glycosylase) and cases wherein the mispair recognition protein merely forms specific complexes with mispairs, which complexes are then recognized by other proteins that modify the DNA in the vicinity of the complex.

Accordingly, the present invention features a method for detecting base pair mismatches in a DNA duplex by utilizing a mismatch recognition protein that forms specific complexes with mispairs, and detecting the resulting DNA-protein complexes by a suitable analytical method.

In addition to methods designed merely to detect base pair mismatches, this invention includes methods for both detecting and localizing base pair mismatches by utilizing components of mismatch repair system.

The present invention also features mispair recognition proteins which have been altered to provide an inherent means for modifying at least one strand of the DNA duplex in the vicinity of the bound mispair recognition protein.

The present invention also concerns systems utilizing an A-G specific mispair recognition protein, for example, the *E. coli* DNA mispair recognition protein that recognizes only A-G mispairs without any apparent requirement for hemimethylation. This protein, the product of the *mutY* gene, is

8

a glycosylase which specifically removes the adenine from an A-G mispair in a DNA duplex. Accordingly, this MutY protein is useful for the specific detection of A-G mispairs according to the practice of the present invention.

The invention also includes the combined use of components of a mismatch repair system along with a recombinase protein. The recombinase protein functions to catalyze the formation of duplex molecules starting with single-stranded molecules obtained from different sources, by a renaturation reaction. Such a recombinase protein is also capable of catalyzing a strand transfer reaction between a single-stranded molecule from one source and double-stranded molecules obtained from a different source. In the presence of a base pair mismatch, formation of duplex regions catalyzed by such a recombinase protein is inhibited by components of a mismatch repair system, e.g., *E. coli* MutS and MutL, proteins. Modulation of recombinase activity by components of a mismatch repair system may involve inhibition of branch migration through regions that generate mismatched base pairs. The combination of a DNA mismatch repair system and a recombinase system provides a very sensitive selection step allowing for the removal of molecules containing a base pair mismatch from a population of newly formed heteroduplex molecules. This procedure provides a selection scheme that can be utilized independent of or in conjunction with the actual mismatch repair reaction.

The invention also features two improvements on the genomic mismatch scanning technique (GMS) of Nelson et al. 4 *Nature Genetics* 11, 1993, used to map regions of genetic identity between populations of DNA molecules.

One improvement provided by the invention features an additional selection step, as described above, for determining genetic variation. The genomic mismatch scanning (GMS) method includes one selection step which is carried out after hybrid formation. The present invention includes an additional step that occurs during hybrid formation, through the use of a protein with recombinase activity along with components of a mismatch repair system. The increase in sensitivity for screening for genetic variation provided by the additional selection step makes possible the use of the GMS technique with larger genomes, e.g., man.

A second improvement provided by the invention features the replacement or modification of the exonuclease III digestion step employed in the GMS method. In the GMS procedure exonuclease III is used to degrade all DNA molecules, except mismatch-free heterohybrids, to molecules containing single-stranded regions, which are subsequently removed. Heterohybrids are duplex molecules which are formed in the method from two molecules which were previously base paired with other molecules (i.e., from different sources). In the instant invention this step is replaced by a procedure that employs all or some of the components of a mismatch repair system. Exo III is a 3' to 5' exonuclease specific for double-stranded DNA, which preferably initiates at blunt or 5' protruding ends. In the GMS procedure DNA molecules are digested with restriction enzymes that produce protruding 3' ends. Although molecules containing protruding 3' ends are not preferred substrates for Exo III, such molecules can be subject to limited attack by the enzyme. Thus, even mismatch-free heterohybrids will be degraded to some extent by Exo III, and will be erroneously removed from the final population of molecules representing those of identity-by-descent. The invention employs components of a mismatch repair system along with dideoxy or biotinylated nucleotide, to avoid the use of Exo III and the potential loss of heterohybrids

5,702,894

9

molecules that are mismatch-free. Homohybrids are digested in the presence of helicase II by *exoVI* RecJ and *exo I*, e.g., natural exonucleases involved in the mismatch repair reaction. The invention also features a modification of the step utilizing *Exo III*, consisting of ligation of duplex DNA molecules at dilute concentrations so as to form closed circular monomer molecules, thus removing any 3' ends which may be subject to degradation by *Exo III*.

The invention includes the use of a mismatch repair system to detect and remove or correct base pair mismatches in a population produced by the process of enzymatic amplification of nucleic acid molecules. DNA polymerase errors that occur during a cycle of enzymatic amplification can result in the presence of mismatched base pair(s) in the population of product molecules. If such errors are perpetuated in subsequent cycles they can impair the value of the final amplified product. The fidelity of the amplification method can be enhanced by including one or more components of a mismatch repair system to either correct the mismatch base pair(s) or to eliminate from the amplified population, molecules that contain mismatch base pair(s). Elimination of molecules containing a base pair mismatch can be accomplished by binding to a protein, such as MutS, or by introduction of a nick in one strand of the duplex so that a full sized product will not be produced in a subsequent round of amplification.

The invention also features methods to remove molecules containing a base pair mismatch through the binding of the mismatch to the components of the mismatch repair system or by the binding of a complex of a mismatch and components of a mismatch repair system to other cellular proteins. Another aspect of the invention for removal of molecules containing a mismatch is through the incorporation of biotin into such a molecule and subsequent removal by binding to avidin.

Another aspect of the invention features use of a mismatch repair system which has a defined 5' to 3' exonuclease function, that is provided by the exonuclease VII or RecJ exonuclease. In other systems a 5' to 3' exonuclease function is provided by exonuclease VII which is present in many preparations of the DNA polymerase III holoenzyme.

The invention also includes kits having components necessary to carry out the methods of the invention.

The mismatch repair systems of the instant invention, e.g., *E. coli*, offer specific and efficient procedures for detection and localization of mismatches and manipulation of DNA containing mismatches that is a reflection of their biological function. All eight possible base pair mismatches are recognized and seven of the eight mismatches are processed and corrected by the system. Although C—C mismatches are not a substrate for repair, MutS does bind weakly to this mispair permitting its detection. In contrast to the electrophoretic migration procedure, the RNase method, or chemical modification procedures, the system does not depend on the destabilization of the DNA helix for detection of mismatches or binding to mismatches. The system features exquisite specificity, and is not subject to non-specific interactions with bases at the ends of linear DNA fragments or non-specific interactions at non-mismatch sites in long molecules.

The detection of fragments containing a mispair is limited only by the intrinsic specificity of the system, for example, detection of better than one G-T mispair per 300 kilobases. Mismatches have been routinely detected with a 6,400 base pair substrate and the system should be applicable to molecules as large as 40–50 kb. This allows for detection of

10

possible single base differences between long DNA sequences, for example, between a complete gene from one individual and the entire genome of another. The invention also enables the localization of any possible single base difference within the sequences of homologous regions of long DNA molecules such as those encoding one or more complete genes and comprising several kbp of DNA.

Several of the methods of the invention result in the covalent alteration of the phosphodiester backbone of DNA molecules. This covalent alteration facilitates analysis of the product DNA molecules especially by electrophoretic methods.

Other features and advantages of the invention will be apparent from the following description of the preferred embodiments thereof, and from the claims.

BRIEF DESCRIPTION OF THE FIGURES

FIG. 1 Heteroduplex substrate for in vitro mismatch correction. The substrate used in some examples is a 6440-bp, covalently closed, circular heteroduplex that is derived from bacteriophage ϕ 1 and contains a single base-base mismatch located within overlapping recognition sites for two restriction endonucleases at position 5632. In the example shown a G-T mismatch resides within overlapping sequences recognized by *Hind III* and *Xho I* endonucleases. Although the presence of the mispair renders this site resistant to cleavage by either endonuclease, repair occurring on the complementary (c) DNA strand yields an A-T base pair and generates a *Hind III*-sensitive site, while correction on the viral (v) strand results in a G-C pair and *Xho I*-sensitivity. The heteroduplexes also contain a single d(GATC) sequence 1024 base pairs from the mismatch (shorter path) at position 216. The state of strand methylation at this site can be controlled, thus permitting evaluation of the effect of DNA methylation on the strand specificity of correction.

FIGS. 2A and 2B Requirement for DNA ligase in mismatch correction. Hemimethylated G-T heteroduplex DNA (FIG. 1, 0.6 μ g, d(GATC) methylation on the complementary DNA strand) was subjected to mismatch repair under reconstituted conditions in a 60 μ l reaction (Table 3, closed circular heteroduplex), or in 20 μ l reactions (0.2 μ g of DNA) lacking MutS protein or ligase, or lacking both activities. A portion of each reaction (0.1 μ g of DNA) was treated with EDTA (10 mM final concentration) and subjected to agarose gel electrophoresis in the presence of ethidium bromide (1.5 μ g/ml; FIG. 2A, lanes 1–4). Positions are indicated for the unreacted, supercoiled substrate (SC), open circles containing a strand break (OC) and covalently closed, relaxed circular molecules (RC). A second sample of each reaction containing 0.1 μ g of DNA was hydrolyzed with *Xho I* and *Cla I* endonucleases (FIG. 1) to score G-T to G-C mismatch correction and subjected to electrophoresis in parallel with the samples described above (FIG. 2B, lanes 5–8). The remainder of the complete reaction (0.4 μ g DNA, corresponding to the sample analyzed in lane 1) was made 10 mM in EDTA, and subjected to electrophoresis as described above. A gel slice containing closed circular, relaxed molecules was excised and the DNA eluted. This sample was cleaved with *Xho I* and *Cla I* and the products analyzed by electrophoresis (FIG. 2B, lane 9).

FIG. 3 Methyl-direction of mismatch correction in the purified system. Repair reactions with the G-T heteroduplex (FIG. 1) were performed as described in Table 3 (closed circular heteroduplex) except that reaction volumes were 20 μ l (0.2 μ g of DNA) and the incubation period was 60

5,702,894

11

minutes. The reactions were heated to 55° for 10 minutes and each was divided into two portions to test strand specificity of repair. G-T to A-T mismatch correction, in which repair occurred on the complementary (c) DNA strand, was scored by cleavage with Hind III and Cla I endonucleases, while hydrolysis with Xho I and Cla I were used to detect G-T to G-C repair occurring on the viral (v) strand. Apart from the samples shown in the left two lanes, all heteroduplexes were identical except for the state of methylation of the single d(GATC) sequence at position 216 (FIG. 1). The state of modification of the two DNA strands at this site is indicated by + and - notation. The G-T heteroduplex used in the experiment shown in the left two lanes (designated 0/0) contains the sequence d(GATT) instead of d(GATC) at position 216, but is otherwise identical in sequence to the other substrates.

FIG. 4 Strand-specific repair of heteroduplexes containing a single strand scission in the absence of Muth protein. Hemimethylated G-T heteroduplex DNAs (FIG. 1, 5 µg) bearing d(GATC) modification on the viral or complementary strand were subjected to site-specific cleavage with near homogeneous Muth protein. Because the Muth-associated endonuclease is extremely weak in the absence of other mismatch repair proteins, cleavage at d(GATC) sites by the purified protein requires a Muth concentration 80 times that used in reconstitution reactions. After removal of Muth by phenol extraction, DNA was ethanol precipitated, collected by centrifugation, dried under vacuum, and resuspended in 10 mM Tris-HCl (pH 7.6), 1 mM EDTA. Mismatch correction of Muth-incised and covalently closed, control heteroduplexes was performed as described in the legend to Table 2 except that ligase and NAD⁺ were omitted. Outside and inside strands of the heteroduplexes depicted here correspond to complementary and viral strands respectively. Values in parentheses indicate repair occurring on the methylated, continuous DNA strand. The absence of Muth protein in preparations of incised heteroduplexes was confirmed in two ways. Preparations of incised molecules were subject to closure by DNA ligase (>80%) demonstrating that Muth protein does not remain tightly bound to incised d(GATC) sites. Further, control experiments in which each Muth-incised heteroduplex was mixed with a closed circular substrate showed that only the open circular form was repaired if Muth protein was omitted from the reaction whereas both substrates were corrected if Muth protein was present (data not shown).

FIG. 5 Requirements for Muth protein and a d(GATC) sequence for correction in the presence of DNA ligase. Hemimethylated G-T heteroduplexes incised on the unmethylated strand at the d(GATC) sequence were prepared as described above in FIG. 4. A G-T heteroduplex devoid of d(GATC) sites (FIG. 4) and containing a single-strand break within the complementary DNA strand at the Hinc II site (position 1) was constructed as described previously (Lahue et al. supra). Mismatch correction assays were performed as described in Table 3, with ligase (20 ng in the presence of 25 µM NAD⁺) and Muth protein (0.26 ng) present as indicated. Table entries correspond to correction occurring on the incised DNA strand, with parenthetical values indicating the extent of repair on the continuous strand. Although not shown, repair of the nicked molecule lacking a d(GATC) sequence (first entry of column 3) was reduced more than an order of magnitude upon omission of MutL, MutS, SSB or DNA polymerase III holoenzyme.

FIG. 6 is a diagrammatic representation of the model system used to evaluate MutS and MutL effects on RecA catalyzed strand transfer.

12

FIG. 7 depicts the effects of MutS and MutL on RecA-catalyzed strand transfer between homologous and quasi-homologous DNA sequences. Solid lines indicate fd—fd strand transfer, while dashed lines correspond to fd—M13 strand transfer. Strand transfer was evaluated in the presence of MutL (solid circles) or MutS (solid squares).

FIG. 8 depicts The MutL potentiation of MutS block to strand transfer in response to mismatched base pairs. Solid lines: fd—fd strand transfer; dashed lines fd—M13 strand transfer; RecA (solid circle); RecA and MutL (solid diamond); RecA and MutS (solid square); RecA, MutL, and MutS (solid triangle).

FIG. 9 depicts the MutS and MutL block of branch migration through regions that generate mismatched base pairs. Solid lines: M13—M13 strand transfer; dashed line fd—M13 strand transfer. RecA only (solid circle and square); RecA, MutS, and MutL (open circle and square).

DESCRIPTION OF PREFERRED EMBODIMENTS

The invention consists of methods utilizing and kits consisting of components of mismatch repair system to detect, and localize DNA base pair mismatches and manipulate molecules containing such mismatches. The invention also features modified mispair recognition proteins and their utilization in the above-mentioned methods and kits. The invention also includes methods and kits comprising components of a mismatch repair system along with proteins with recombinase activity. The invention also consists of methods to improve the GMS technique to detect regions of homology-by-descent.

Methods for Detecting the Presence and Localization of Mismatched Base Pairs by Complex Formation with a Mismatch Recognition Protein

One embodiment of the invention features a diagnostic method for detecting a base pair mismatch in a DNA duplex. The method comprises the steps of contacting at least one strand of a first DNA molecule with the complementary strand of a second DNA molecule under conditions such that base pairing occurs, contacting a DNA duplex potentially containing a base pair mismatch with a mispair recognition protein under conditions suitable for the protein to form a specific complex only with the DNA duplex having a base pair mismatch, and not with a DNA duplex lacking a base pair mismatch, and detecting the complex as a measure of the presence of a base pair mismatch in the DNA duplex.

By "mismatch" is meant an incorrect pairing between the bases of two nucleotides located on complementary strands of DNA, i.e., bases pairs that are not A:T or G:C.

In the practice of this method, the two DNA's or two DNA samples to be compared may comprise natural or synthetic sequences encoding up to the entire genome of an organism, including man, which can be prepared by well known procedures. Detection of base sequence differences according to this method of this invention does not require cleavage (by a restriction nuclease, for example) of either of the two DNAs, although it is well known in the art that rate of base pair formation between complementary single-stranded DNA fragments is inversely related to their size. This detection method requires that base sequence differences, to be detected as base pair mismatches lie within a region of homology constituting at least about 14 consecutive base pairs of homology between the two DNA molecules, which is about the minimum number of base pairs generally required to form a stable DNA duplex. Either one or both of the strands of the first DNA may be selected for examination, while at least one strand of the second DNA

5,702,894

13

complementary to a selected first DNA strand must be used. The DNA strands, particularly those of the second DNA, advantageously may be radioactively labeled to facilitate direct detection, according to procedures well known in the art.

By "mismatch recognition protein" is meant a protein of a mismatch repair system that specifically recognizes and binds to a base pair mismatch, e.g., *E. coli* MutS.

Methods and conditions for contacting the DNA strands of the two DNAs under conditions such that base pairing occurs are also widely known in the art.

In preferred embodiments of this aspect of this invention, the mismatch recognition protein is the product of the *mutS* gene of *E. coli*, or species variations thereof, or portions thereof encoding the recognition domain. The protein recognizes all eight possible base pair mismatches, detection of the DNA:protein complex comprises contacting the complexes with a selectively absorbent agent under conditions such that the protein:DNA complexes are retained on the agent while DNA not complexed with protein is not retained and measuring the amount of DNA in the retained complexes, the absorbent agent is a membranous nitrocellulose filter, detection of the DNA:protein complex further includes the step wherein an antibody specific for the base mismatch recognition protein is employed, the base mismatch recognition protein is the product of the *mutS* gene of *S. typhimurium* the *hexA* gene of *S. pneumoniae* or the *MSH1* and *MSH2* genes of yeast, and wherein the step for detecting the DNA:protein complex further includes a step wherein the electrophoretic mobility of the DNA:protein complex is compared to uncomplexed DNA.

The ability of the MutS protein to recognize examples of all eight single base pair mismatches within double-stranded DNA, even including C—C mismatches which do not appear to be corrected in vivo, is demonstrated by the fact that MutS protein protects DNA regions containing each mismatch from hydrolysis by DNase I (i.e., by "Dnase I footprint" analyses), as recently reported (Su, S.-S., et al., 1988, *J. Biol. Chem.*, 263, 6829–6835). The affinity of MutS protein for the different mismatches that have been tested varies considerably. Local sequence environment may also affect the affinity of the MutS protein for any given base mismatch; in other words, for example, the affinity for two specific cases of A—C mismatches, which are surrounded by different sequences, may not be the same. Nevertheless, no examples of base mismatches have been found that are not recognized by isolated MutS protein. Accordingly, this method of the invention detects all mismatched base pairs.

It should be particularly noted that the DNA duplexes which MutS recognizes are not required to contain GATC sequences and, hence, they do not require hemimethylation of A's in GATC sequences, the specific signal for the full process of methyl-directed mismatch correction in vivo; therefore, use of MutS in this method allows recognition of a DNA base mismatch in DNAs lacking such methylation, for instance, DNAs isolated from human tissues.

By "species variation" is meant a protein which appears to be functionally and in part, at least, structurally homologous to the *E. coli* MutS protein. One example of such a protein has also been discovered in a methyl-directed mismatch correction system in *Salmonella typhimurium* bacteria (Pang et al., 1985, *J. Bacteriol.*, 163, 1007–1015). The gene for this protein has been shown to complement *E. coli* strains with mutations inactivating the *mutS* gene and the amino acid sequence of its product shows homology with that of the *E. coli* MutS protein. Accordingly, this *S. typhimurium* protein is also suitable for the practice of this aspect of the

14

present invention. Other organisms, including man, are known to possess various systems for recognition and repair of DNA mismatches, which, as one skilled in the art would appreciate, comprise mismatch recognition proteins functionally homologous to the MutS protein. Nuclear extracts derived from HeLa and *Drosophila melanogaster* K₅₆₂ cell lines has been shown to support efficient strand specific mismatch correction in vitro (Holmes et al., 1990, *Proc. Natl. Acad. Sci. USA* 87, 5837–5841, which is incorporated herein by reference), and this reaction has been shown to occur by a mechanism similar to that of the bacterial reaction (Fany and Modrich 268 *J. Biol. Chem.* 11838, 1993). Furthermore, genes encoding proteins that are homologous to bacterial MutS at the amino acid sequence level have been demonstrated in human (Fujii and Shimada 264 *J. Biol. Chem.* 10057, 1989) and yeast (Reenan and Kolodner 132 *Genetics* 963, 1992) and *S. pneumoniae* (Priebe et al., 1970 *J. Bacteriol.* 190, 1988). Accordingly, it is believed that such DNA base mismatch recognition proteins may also be suitable for use in the present invention.

By "protein encoding the recognition domain" is meant a region of the mismatch recognition protein which is involved in mismatch recognition and binding. Such a domain comprises less than the complete mismatch recognition protein.

By a "selectively adsorbent agent" is meant any solid substrate to which protein:DNA complexes are retained on the agent while DNA not complexed with protein is not retained, such agents are known to those skilled in the art. Absent radioactive labeling of at least one strand used to form the DNA duplexes, the DNA in complexes on the filter may be detected by any of the usual means in the art for detection of DNA on a solid substrate, including annealing with complementary strands of radioactive DNA.

The nitrocellulose filter method for detecting complexes of MutS protein with base mismatches in DNA has been reported in detail (Jiricny, J. et al., 1988, *Nuc. Acids Res.* 16, 7843–7853, which is hereby incorporated herein by reference). Besides simplicity, a major advantage of this method for detecting the DNA:protein complex over other suitable methods is the practical lack of a limitation on the size of DNA molecules that can be detected in DNA:protein duplexes. Therefore, this embodiment of this method is in principle useful for detecting single base sequence differences between DNA fragments as large as can be practically handled without shearing.

By "electrophoretic mobility" is meant a method of separating the DNA:protein complexes from DNA that does not form such complexes on the basis of migration in a gel medium under the influence of an electric field. DNA:protein complexes are less mobile than naked DNA. Such methods based on electrophoretic mobility are known to those skilled in the art. The DNA in the DNA:protein complexes may be detected by any of the usual standard means for detection of DNA in gel electrophoresis, including staining with dyes or annealing with complementary strands of radioactive DNA. Detecting complexes comprising the MutS base mismatch recognition protein and mismatches in DNA duplexes is also described in the foregoing reference (Jiricny, J. et al., 1988, *Nuc. Acids Res.* 16, 7843–7853). Under the usual conditions employed in the art for detecting specific DNA:protein complexes by gel electrophoresis, complex formation of a protein with a double-stranded DNA fragment of up to several hundred base pairs is known to produce distinguishable mobility differences.

Antibodies specific for a DNA mismatch recognition protein can be prepared by standard immunological techniques known to those skilled in the art.

5,702,894

15

Other suitable analytical methods for detecting the DNA protein complex include immunodetection methods using an antibody specific for the base mismatch recognition protein. For example, antibodies specific for the *E. coli* MutS protein have been prepared. Accordingly, one immunodetection method for complexes of MutS protein with DNA comprises the steps of separating the DNA:protein complexes from DNA that does not form such complexes by immunoprecipitation with an antibody specific for MutS protein, and detecting the DNA in the precipitate. According to the practice of this aspect of the invention, quantitative immunoassay methods known in the art may be employed to determine the number of single base mismatches in homologous regions of two DNA molecules, based upon calibration curves that can be established using complexes of a given mismatch recognition protein with DNA duplexes having known numbers of mismatches.

Another aspect of the invention features a method for detecting and localizing a base pair mismatch in a DNA duplex. The method includes contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting double-stranded DNA duplexes with a mismatch recognition protein under conditions such that the protein forms specific complexes with mismatches, subjecting the duplex molecules to hydrolysis with an exonuclease under conditions such that the complex blocks hydrolysis, and determining the location of the block to hydrolysis by a suitable analytical method.

"Hydrolysis with an exonuclease" is a procedure known to those skilled in the art and utilizes enzymes possessing double-strand specific exonuclease activity, e.g., *E. coli* exonuclease III, RecBCD exonuclease, lambda exonuclease, and T7 gene 6 exonuclease.

By "block to hydrolysis" is meant interference of hydrolysis by the exonuclease. Such protection can result from the mismatch recognition protein protecting the DNA to which it is bound.

By "suitable analytical method" is meant any method that allows detection of the block to exonuclease digestion, such analysis of molecules by gel electrophoresis. Such methods are known to those skilled in the art. Methods for Detecting and Localizing Base Pair Mismatches by Mismatch Repair System Strand Modification Reactions

In addition to methods that detect base sequence differences, this invention provides methods for both detecting and localizing a base pair mismatch in a DNA duplex. One method includes contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting double-stranded DNA duplexes with a mismatch recognition protein under conditions such that the protein forms specific complexes with mismatches and thereby directs modification of at least one strand of the DNA in the resulting DNA:protein complexes in the vicinity of the DNA:protein complex, and determination of the location of the resulting DNA modification by a suitable analytical method.

By "modification" is meant any alteration for which there is a means of detection, for instance a chemical modification including breaking of a chemical bond resulting in, as examples, cleavage between nucleotides of at least one DNA strand or removal of a base from the sugar residue of a nucleotide. Specific means for modifying DNAs in the vicinity of the DNA:protein complex are provided below for several embodiments of this aspect of the invention, together

16

with interpretations of the phrase "in the vicinity of", as appropriate to the practical limitations of the modification approach in each instance.

Suitable analytical methods for determining the location of the modification are known to those skilled in the art. Such a determination involves comparison of the modified DNA molecule with the homologous unmodified DNA molecule.

In preferred embodiments of this aspect of the invention, the mismatch recognition protein is the product of the mutS gene of *E. coli* or another functionally homologous protein; the step in which the DNA is modified in the vicinity of the DNA:protein complex further comprises contacting the DNA:MutS protein complex with a defined set or subset of *E. coli* DNA mismatch repair proteins (comprising *E. coli* MutH, MutL, DNA helicase II, single-stranded DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, and exonuclease VII (or RecJ exonuclease), or species variations of these activities), ATP and one or more dideoxynucleoside-5'-triphosphates or in the absence of exogenous deoxyribonucleoside-5'-triphosphate under conditions that produce a discontinuity in one or both strands of the DNA duplex in the vicinity of the mismatch.

DNA used in such an analysis are to be unmethylated or hemimethylated at on the 6-position of the adenine base in GATC sequences. With the exception of DNAs from some bacterial species, the chromosomes of most organisms naturally lack this modification. In those cases where hemimethylation of otherwise GATC unmodified molecules is desired, this can be accomplished by use of *E. coli* Dam methylase as is well known in the art. Symmetrically methylated DNA prepared by use of this enzyme is denatured and subsequently reannealed with single-stranded sequences representing an homologous (or largely so) DNA. If necessary, hemimodified molecules produced by this reannealing procedure can be separated from unmethylated is symmetrically methylated duplexes which can also result from the annealing procedure. As is well known in the art, this can be accomplished by subjecting annealed products to cleavage by DpnI and MboI endonucleases. The former activity cleaves symmetrically methylated duplex DNA at GATC sites while unmodified duplex DNA is subject to double strand cleavage only at unmodified GATC sites by the latter activity. Since hemimodified DNA is resistant to double strand cleavage by both DpnI and MboI, desired hemimethylated products can be separated on the basis of size from the smaller fragments produced by DpnI and MboI cleavage, for example by electrophoretic methods.

By "discontinuity in one or both strands of the DNA duplex" is meant a region which consists of a break in the phosphodiester backbone in one or both strands, or a single-stranded gap in a duplex molecule.

One aspect of this preferred embodiment involves contacting the DNA:MutS protein complex with *E. coli* MutL and MutH proteins (or species variations thereof) in the presence of ATP and an appropriate divalent cation cofactor (e.g., Mg^{2+}) so that mismatch-containing molecules will be subject to incision at one or more GATC sites in the vicinity of the mismatch. Such incision events can be monitored by a suitable analytical method for size detection such as electrophoresis under denaturing condition.

A second aspect of this preferred embodiment involves contacting the DNA:MutS complex with a defined *E. coli* mismatch correction system consisting of *E. coli* MutH, MutL, DNA helicase II, single-stranded DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, and exonuclease VII (or RecJ exonuclease), or species

5,702,894

17

variants of these activities. ATP in the absence of exogenous deoxyribonucleoside-5'-triphosphates or in the presence of one or more dideoxynucleoside-5'-triphosphates such that single-stranded gaps are produced in the vicinity of the complexed protein; the method for determining the location of the single-stranded gaps with the DNA duplex further includes analysis of electrophoretic mobility of treated samples under denaturing conditions of the steps of cleaving the DNA with a single-stranded specific endonuclease, and comparing the electrophoretic mobilities of the cleaved fragments with unmodified DNA fragments under non-denaturing conditions; the step for modifying the DNA duplex in the vicinity of the complexed protein comprises contacting the complexes with proteins of a mismatch repair system, ATP and a divalent cation under conditions such that an endonucleolytic incision is introduced at one or more GATC sequences in the duplex molecule.

An example of a complete defined mismatch correction system comprises the following purified components: *E. coli* MutH, MutL, and MutS proteins, DNA helicase II, single-stranded DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, DNAligase, ATP, and the four deoxynucleoside-5'-triphosphates. This set of proteins can process seven of the eight base-base mismatches in a strand-specific reaction that is directed by the state of methylation of a single GATC sequence located 1 kilobase from the mispair. This defined system is described further in Example 1, below. The 5' to 3' exonuclease function can either be supplied by either DNA polymerase III holoenzyme preparations that contain this activity or as a separate defined component consisting of exonuclease VII or RecJ exonuclease. It should be noted that the lack of ability to repair C—C base mispairs in this embodiment of this aspect of the present invention is not a major limitation of the method for detecting all possible base sequence differences between any two naturally occurring DNA sequences because mutations that would give rise to a C—C mispair upon hybridization would also give rise to a G—G mismatch when the complementary strands are hybridized.

For the purpose of generating single-stranded gaps in the vicinity of the DNA:MutS protein complexes, DNA duplexes containing mispaired base pairs are contacted with the defined mismatch correction system under the standard conditions described in Example 1, Table 3 (Complete reaction), except for the following differences:

- (i) exogenous dNTPs are omitted; or (ii) 2', 3'-dideoxynucleoside-5'-triphosphates (ddNTPs) at suitable concentrations (10 to 100 μ M) are substituted for dNTPs; or (iii) reactions containing dNTPs are supplemented with ddNTPs at a suitable concentration to yield a chain termination frequency sufficient to inhibit repair of single-strand gaps. In cases (i)–(iii) DNA ligase may be omitted from the reaction. In cases (ii) and (iii) all four ddNTPs may be present; however, it is expected that the presence of one, two, or three ddNTPs will prove sufficient to stabilize single strand gaps via chain termination events. While it is expected that most applications of these gap forming protocols will utilize MutH, it is pertinent to note that the requirement of methyl-directed strand incision by MutH may be obviated by provision of a single-strand nick by some other means within the vicinity of the mispair, as described in Example 1, FIG. 5. A suitable means for inducing such nicks in DNA is limited contact with a nuclease, Dnase I, for example; under conditions that are well known in the art, this approach creates nicks randomly throughout double-stranded DNA molecules at suitable intervals for allowing the mispair correction

18

system to create single-stranded gaps in the vicinity of a mispair anywhere in the DNA.

It should be noted that in this embodiment of this method for localizing mismatch base pairs, "in the vicinity of" a base mispair is defined practically by the size of the single-strand gaps typically observed under above conditions, namely up to about one kbp from the mismatched base pair.

By "determining the location of the single-stranded gaps within the DNA duplex" entails the steps of:

- (i) Cleaving the DNA with at least one restriction endonuclease (either prior or subsequent to contact of the preparation with mismatch repair activities) followed by comparison of electrophoretic mobilities under denaturing conditions of the resulting modified DNA fragments with DNA restriction fragments not contacted with the defined mismatch correction system; or (ii) Cleaving the DNA with at least one restriction endonuclease and with a single-strand specific endonuclease, followed by comparison of the electrophoretic mobilities under native conditions of the resulting modified DNA fragments with DNA restriction fragments not contacted with the defined mismatch correction system. Suitable single-strand specific endonucleases include the S1 single-stranded specific nuclease, for example, or other functionally similar nucleases well known in the art. In the cases of either (i) or (ii), additional restriction mapping may be performed as needed to further localize any fragment modifications observed in initial application of the method, until, if desired, a restriction fragment of convenient size for direct sequence determination is obtained for direct comparisons of sequences of the two DNA molecules in the vicinity of the base sequence difference.

By "proteins of a mismatch repair system" are meant a protein that contains a GATC endonuclease, a mispair recognition protein, and proteins that participate in the activation of the GATC endonuclease.

By "divalent cation" is meant a cofactor for the GATC endonucleases, e.g., $MgCl_2$.

By "endonucleolytic incision" is meant cleavage of a DNA fragment containing a mismatched base pair at unmethylated or hemimethylated GATC sequences in the vicinity of a mismatch.

"Size fractionation by electrophoretic mobility under denaturing conditions" is a procedure well known by those skilled in the art. Gel Electrophoresis can either be conventional or pulse-field.

Modification of Mismatch Recognition Proteins and Uses

The present invention also includes forms of mispair recognition proteins which have been altered to provide means for modifying at least one strand of the DNA duplex in the vicinity of the bound mispair recognition protein.

In preferred embodiments of this aspect of the invention, the altered mispair recognition protein is the modified product of the mutS gene of *E. coli* or is another functionally homologous modified protein to which is attached an hydroxyl radical cleaving function; the altered mispair recognition protein may comprise only a segment of the native molecule containing the mispair recognition domain; the hydroxyl radical cleaving function is selected from the group consisting of the altered mispair recognition protein wherein the hydroxyl radical cleaving function is selected from the group consisting of the 1,10-phenanthroline-copper complex, the EDTA iron complex, and the copper binding domain of serum albumin; the altered mispair recognition protein is the product of the mutS gene of *E. coli* or of another functionally homologous protein to which is attached attachment a DNA endonuclease activity capable of

5,702,894

19

cleaving double-stranded DNA; the endonuclease activity is provided by the DNA cleavage domain of FokI endonuclease.

By "altered mispair recognition protein" is meant a mispair recognition protein that not only recognizes and binds to a base pair mismatch, but possess the ability to modify a strand of a DNA molecule containing such a mismatch.

Several methods for attaching an hydroxyl radical cleaving function to a DNA binding protein are known in the art. For example, lysyl residues may be modified by chemically attaching the 1,10-phenanthroline-copper complex to lysine residues, resulting in conversion of a DNA binding protein into a highly efficient site-specific nuclease that cleaved both DNA strands (in the presence of hydrogen peroxide as a coreactant) within the 20 base pair binding site of the protein, as determined by DNase I footprinting (C.-H. Chen and D. S. Sigman, 1987, *Science*, 237, 1197). Chemical attachment of an EDTA-iron complex to the amino terminus of another DNA binding protein similarly produced a sequence specific DNA cleaving protein that cut both strands of the target DNA within a few bases of recognition site of similar size (J. P. Sluka, et al., 1987, *Science*, 235, 777).

An alternate means for attaching the hydroxyl radical cleaving function to this same protein involved extension of the amino terminus with the three amino acids, Gly-Gly-His, which is consensus sequence for the copper-binding domain of serum albumin (D. P. Hack et al., 1988, *J. Am. Chem. Soc.*, 110, 7572-7574). This approach allows for preparation of such an artificial DNA cleaving protein directly by recombinant methods, or by direct synthesis using standard solid phase methods, when the peptide is sufficiently short as it was in this case (55 residues including the 3 added amino acids), thereby avoiding the need for an additional chemical modification step of the reagent which is both time consuming and difficult in large scale production. In contrast to the EDTA-iron complex, the particular peptide sequence constructed in this instance cleaved only one example out of four recognition sites in different sequence environments.

Nevertheless, one skilled in the art of protein engineering would appreciate that this general approach for converting a DNA binding protein into a DNA cleaving protein by attachment of an hydrogen radical cleavage function is widely applicable. Hence, DNA base mispair recognition proteins which normally only bind to DNA are modified to cleave DNA by attachment of an hydroxyl radical cleavage function, according to the practice of this aspect of this invention, without undue experimentation, by adjustment of appropriate variables taught in the art, particularly the chemical nature and length of the "spacer" between the protein and the metal binding site.

Additional altered forms of mispair recognition proteins that modify at least one strand of the DNA in a DNA:protein complex in the vicinity of the bound protein according to the present invention include proteins comprising the portions or "domains" of the unmodified base mispair recognition enzymes that are essential for binding to a DNA mispair. These essential DNA binding domains further comprise peptide sequences that are most highly conserved during evolution; such conserved domains are evident, for example, in comparisons of the sequences of the *E. coli* MutS protein with functionally homologous proteins in *S. typhimurium* and other structurally similar proteins. Accordingly, peptide sequences of a DNA base mispair recognition protein that are protected from proteases by formation of specific complexes with mispairs in DNA and, in addition or in the alternative, are evolutionarily conserved, form the basis for a particularly preferred embodiment of this aspect of the

20

present invention, since such peptides constitute less than half the mass of the intact protein and, therefore, are advantageous for production and, if necessary, for chemical modification to attach a cleavage function for conversion of the DNA binding protein into a DNA cleavage protein specific for sites of DNA base mispairs.

The DNA cleavage domain of FokI endonuclease has been defined (Li et al, 1992, *Proc. Natl. Acad. Sci. U.S.A.*, 89:4275).

Another embodiment of this aspect of the invention consists of a method for detecting and localizing a base pair mismatch within a DNA duplex, including the steps of contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs; contacting resultant duplex DNA molecules with an altered mispair recognition protein, under conditions such that the protein forms specific complexes with a mispair and thereby directs modification of at least one strand of the DNA in the resulting DNA protein complexes in the vicinity of the DNA protein complex, and determining the location of the modification of the DNA by a suitable analytic method.

In the detection and localization of a base pair mismatch method according to this embodiment which employs an altered mispair recognition protein, and the modification comprises double-stranded cleavage of the DNA within the vicinity of any base mispair wherein the "vicinity" substantially corresponds to the sequence of DNA protected by the binding of the protein to a base mispair, generally within about 20 base pairs. A single-strand specific nuclease, S1, for instance, may be used to augment cleavage by the modified base mispair recognition protein in the event that a single-strand bias is suspected in the cleavage of any DNAs with which the protein forms a specific complex. Alternatively, DNA's subject to cleavage by the modified mispair recognition protein may be analyzed by electrophoresis under denaturing conditions. Location of the modification is by suitable analytical methods known to those skilled in the art. Methods Utilizing Mismatch Repair Systems to Detect A-G Base Pair Mismatches

In a preferred embodiment, a method for detecting and localizing A-G mispairs in a DNA duplex, includes the steps of contacting at least one strand of the first DNA molecule with the complementary strand of the second DNA molecule under conditions such that base pairing occurs; contacting resultant duplex DNA molecules with a mispair recognition protein that recognizes A-G mispairs and an apurinic endonuclease or lyase under conditions such that in the presence of a mismatch an endonucleolytic incision is introduced in the duplex molecule, and determining the location of the incision by a suitable analytic method.

In preferred embodiments the A-G mispair recognition protein is the product of the mutY gene of *E. coli*; and the analytical method includes gel electrophoresis.

The present invention also comprises DNA mispair recognition protein that recognizes primarily A-G mispairs without any apparent requirement for hemimethylation. One example of this protein is the product of the mutY gene of *E. coli*, is a glycosylase which specifically removes the adenine from an A-G mispair in a DNA duplex. The MutY protein has been purified to near homogeneity by virtue of its ability to restore A-G to C-G mismatch correction to cell-free extracts (K. G. Au et al., *Proc. Nat. Acad. Sci. U.S.A.*, 85, 9163, 1988) of a mutS mutY double mutant strain of *E. coli*, as described in Example 2, below. Its electrophoretic migration in the presence of dodecyl sulfate is consistent with a molecular weight of 36 kDa, and it

5,702,894

21

apparently exists as a monomer in solution. MutY, an apurinic (AP) endonuclease, DNA polymerase I, and DNA ligase are sufficient to reconstitute MutY-dependent, A-G to C*G repair in vitro. A DNA strand that has been depurinated thusly by the MutY protein is susceptible to cleavage by any of several types of AP endonuclease or lyase (e.g. human AP endonuclease II) or by piperidine, under conditions that are well known in the art. The cleavage products are then analyzed by gel electrophoresis under denaturing conditions. Accordingly, this MutY protein is useful in a method for the specific detection and localization of A-G mispairs, according to the practice of the present invention, and hence identification of A*T to C*G or G*C to T*A mutations. Sources of DNA Fragments to be Analyzed

In another embodiment of the invention, DNA molecules are obtained from the following sources: different individuals of the same species, individuals of different species, individuals of different kingdoms, different tissue types, the same tissue type in different states of growth, different cell types, cells of the same type in different states of growth, and cells of the same origin in different stages of development, and cells of the same type that may have undergone differential somatic mutagenesis, e.g., one class of which may harbor per-cancerous mutation(s).

In a preferred embodiment, the DNA molecules comprise a probe sequence that has been at least partially characterized.

By "probe sequence that has been at least partially characterized" is meant a DNA molecule from any source that has been characterized by restriction mapping or sequence analysis, such techniques are known to those skilled in the art.

Kits Comprising a Mismatch Recognition Protein

Another aspect of the invention features assay kits designed to provide components to practice the methods of the invention.

In one aspect the invention features an assay kit for detecting a base pair mismatch in a DNA duplex. The kit comprises one or more of the following components: an aliquot of a mismatch recognition protein, an aliquot of control oligonucleotides, and an exonuclease.

In a preferred embodiment the mismatch recognition protein is the product of the mutS gene of *E. coli*.

By "control oligonucleotides" is meant oligonucleotides for assaying the binding of the mismatch repair protein to a base pair mismatch. One set of oligonucleotides are perfectly homologous (negative control) and thus are not bound by the mismatch recognition protein. Another set of oligonucleotides containing a base pair mismatch (positive control) and thus are bound by the mismatch recognition protein.

By "exonuclease" is meant enzymes possessing double-strand specific exonuclease activity, e.g., *E. coli* exonuclease III, RecBCD exonuclease, lambda exonuclease, and T7 gene 6 exonuclease.

Another aspect of the invention features an assay kit for detecting and localizing a base pair mismatch in a DNA duplex. The kit comprises one or more of the following components: an aliquot of all or part of a mismatch repair system, an aliquot of dideoxynucleoside triphosphates; and a single-strand specific endonuclease.

By "all or part of a mismatch repair system" is meant either the complete system which is capable of repairing a base pair mismatch, for example, the three *E. coli* proteins MutH, MutL, and MutS, DNA helicase II, single-strand binding protein, DNA polymerase III, exonuclease I, exonuclease VII or RecJ exonuclease, DNA ligase and ATP, or

22

only the three proteins MutH, MutL, and MutS, along with ATP such that an endonucleolytic incision is made at a GATC site, with no subsequent repair reaction taking place.

In preferred embodiments the mismatch repair system includes: the products of the *E. coli* mutH, mutL, and mutS genes, or species variations thereof, DNA helicase II, single-strand DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, exonuclease VII or RecJ exonuclease, DNA ligase, and ATP, the mismatch repair system includes only the products of the *E. coli* mutH, mutL, and mutS genes, or species variations thereof, and ATP.

Another embodiment of the invention features an assay kit for detecting and localizing a base pair mismatch in a DNA duplex comprising an aliquot of a modified mismatch recognition protein.

In a preferred embodiment the mismatch recognition protein is the product of the mutS gene of *E. coli*.

A further embodiment of this aspect of the invention features an assay kit for detecting and localizing an A-G mismatch within a DNA duplex. The kit comprises one or more of the following components: an aliquot of an A-G mismatch recognition protein; and an aliquot of an apurinic endonuclease or lyase.

In a preferred embodiment the A-G mismatch recognition protein is the product of the MutY gene of *E. coli*.
Methods Utilizing Mismatch Repair Systems and Recombinase Proteins

In a further aspect, the invention features a method for eliminating DNA molecules containing one or more mismatches from a population of heterohybrid duplex DNA molecules formed by base pairing of single-stranded DNA molecules obtained from a first source and a second source. The method includes digesting genomic DNA from the first and the second source with a restriction endonuclease, methylating the DNA of one of the sources, denaturing the DNA from one or both sources, mixing the DNA molecules from the first and the second source in the presence of a recombinase protein, proteins of a mismatch repair system that modulate the recombinase protein, single-strand binding protein, and ATP under conditions such that DNA duplexes form in homologous regions of the DNA molecules from the first and the second source and the presence of a base pair mismatch results in regions that remain single-stranded, and removing molecules that contain single-stranded regions from the population.

By "heterohybrid" is meant a duplex DNA molecule that consists of base-paired strands originating from two different sources, such that one strand of the duplex is from one source (first source) and the other strand is from another source (second source).

The "source" of DNA molecules designates the origin of the genomic DNA used in the method. The first and second sources are different, i.e., not from the same cell of the same individual.

By "restriction endonuclease" is meant an enzyme which recognizes specific sequences in double-stranded DNA and introduces breaks the phosphodiester backbone of both strands. For use in the current invention restriction endonucleases that digest genomic DNA or cDNA into fragments of approximately 4 to 20 kilobases are preferred.

By "methylating" is meant the process by which a methyl group is attached to the adenine residue of the sequence "GATC". This reaction is carried by enzymes well known in the art, such as the DAM system of *E. coli*.

By "denaturing" is meant the process by which strands of duplex DNA molecules are no longer base paired by hydrogen bonding and are separated into single-stranded

5,702,894

23

molecules. Methods of denaturation are well known to those skilled in the art and include thermal denaturation and alkaline denaturation.

By "recombinase protein" is meant a protein that catalyzes the formation of DNA duplex molecules. Such a molecule is capable of catalyzing the formation of duplex DNA molecules from complementary single-stranded molecules by renaturation or by catalyzing a strand transfer reaction between a single-stranded molecule and a double-stranded molecule. Examples of such a protein are the RecA proteins of *E. coli* and *S. typhimurium*.

By "proteins of a mismatch repair system that modulate the recombinase protein" are meant components of a system which recognizes and corrects base pairing errors in duplex DNA molecules and also influence the activity of a recombinase protein. For example, a mismatch recognition protein, e.g., MutS, and a protein that interacts with the mismatch repair protein, e.g., MutL, together inhibit duplex formation catalyzed by the recombinase protein in the presence of a base pair mismatch. Such modulation of the recombinase protein results in single-stranded regions downstream of the base pair mismatch.

In preferred embodiments, the recombinase protein is the *E. coli* RecA protein, the mismatch repair system is from *E. coli* and the components are the MutS and MutL proteins, the sources of DNA are different individuals of the same species, individuals of different species, individuals of different kingdoms, different tissue types, the same tissue type in different states of growth, different cell types, cells of the same type in different states of growth, cells of the same origin in different stages of development, and cells of the same origin that may have undergone differential somatic mutagenesis, the method of removing molecules containing single-stranded regions is by chromatography on benzoylated naphthoylated DEAE, the method of removing molecules containing single-stranded regions is by treatment with a single-strand specific nuclease.

The MutS, MutL protein, along with single-strand binding protein and ATP are involved in modulation of the *E. coli* RecA protein in catalyzing heteroduplex formation.

The method for removing molecules containing single-strands from double-stranded molecules by the use of chromatography with benzoylated naphthoylated DEAE is well known to those skilled in the art.

By "single strand specific nuclease" is meant an enzyme that specifically degrades single-stranded regions of DNA molecules and do not degrade double stranded regions. Examples of such nucleases are: S1, mung bean, T7 gene 3 endonuclease and P1 nuclease.

In another aspect, the invention features a method for eliminating DNA molecules containing one or more mismatches from a population of heterohybrid duplex DNA molecules formed by a strand transfer reaction between duplex DNA molecules obtained from a first source and denatured DNA molecules from a second source. The method includes digesting genomic DNA from the first and the second source with a restriction endonuclease, methylating the DNA of one of the sources, denaturing the DNA from the second source, mixing the DNA molecules from the first and the second source in the presence of a protein which catalyzes strand transfer reactions, proteins of a mismatch repair system that modulate the protein with strand transfer activity, single strand binding protein, and ATP under conditions such that DNA heteroduplexes form in homologous regions of the DNA molecules from the first and the second source by strand transfer reaction and the presence of a base pair mismatch results in regions that remain single-stranded,

24

and removing molecules that contain the single-stranded regions from the population.

By "strand transfer reaction is meant" a three strand reaction between duplex DNA from one source and single-stranded DNA from another source in which one strand of the duplex is displaced by a single-stranded molecule.

By "a protein which catalyzes strand transfer reaction" is meant proteins such as: RecA, homologs of RecA, and proteins with branch migration enhancing activities such as RuvA, RuvB, RecG.

In preferred embodiments, the strand transferase protein is the *E. coli* RecA protein, the mismatch repair system is from *E. coli* and the components are the MutS and MutL proteins, the sources are different individuals of the same species, individuals of different species, individuals of different kingdoms, different tissue types, the same tissue type in different states of growth, different cell types, cells of the same type in different states of growth, and cells of the same origin in different stages of development, cells of the same origin that may have undergone differential somatic mutagenesis (e.g., normal as opposed to pre-tumor cells), a probe sequence that has been at least partially characterized, the method of removing molecules containing single-stranded regions is by chromatography on benzoylated naphthoylated DEAE, the method of removing molecules containing single-stranded regions is by treatment with a single strand specific nuclease.

Methods of Improving the Genomic Mismatch Scanning Technique

In another aspect the invention features the utilization of a recombinase or strand transferase and proteins of a mismatch repair system that modulate the recombinase or strand transferase, in the hybridization step of the genomic mismatch scanning technique. Formation of duplex molecules catalyzed by a recombinase or strand transferase protein which is modulated by components of a mismatch repair system, provide an additional selection step in the GMS method.

By "genomic mismatch scanning" is meant a technique to identify regions of genetic identity between two related individuals. Such a technique has been described by Nelson et al, 4 *Nature Genetics* 11, 1993.

In a further embodiment the invention features a method of genomic mismatch scanning such that heterohybrid DNA molecules containing a base pair mismatch are removed, without the use of exonuclease III. The method comprises the steps of contacting a population of heterohybrid DNA molecules potentially containing base pair mismatches with all the components of a DNA mismatch repair system in the absence of dNTP's or in the presence of one or more dideoxy nucleoside triphosphates under conditions such that single-stranded gaps are generated in DNA fragments that contained a base pair mismatch and removing the molecules containing single-stranded gaps.

In preferred embodiments the DNA mismatch repair system is the *E. coli* methyl-directed mismatch repair system; removal of molecules containing single-stranded regions is by chromatography on benzoylated naphthoylated DEAE; removal of molecules containing single-stranded regions is by treatment with a single-strand specific nuclease.

In a further embodiment, the invention features another variation of the method of genomic mismatch scanning such that heterohybrid DNA molecules containing base pair mismatches are removed, without the use of exonuclease III. The method comprises the steps of contacting a population of heterohybrid DNA molecules potentially containing base

5,702,894

25

pair mismatches with all the components of a DNA mismatch repair system and biotinylated nucleoside triphosphates under conditions such that biotinylated nucleotides are incorporated into DNA fragments that contained a base pair mismatch and, removing the molecules containing biotinylated molecules by binding to avidin.

Substitution with biotinylated nucleotides and binding of molecules that have incorporated these nucleotides are procedures well known to those skilled in the art. This procedure allows fractionation of a population of hybrid DNA molecules into two fractions: (i) A mismatch free fraction which fails to adhere to avidin; and (ii) A population that originally contained mispairs and which binds to avidin. The former can be utilized in the GMS procedure. The latter, avidin-bound class can be employed for other purposes. For example, when prepared using heterohybrid DNA produced by annealing DNA from two related haploid organisms the biotinylated sequences correspond to those DNA regions that vary genetically between the two organisms. Such sequences can thus be applied to determination of the molecular basis of genetic variation of organisms in question, e.g., pathogenic versus nonpathogenic microbial subspecies.

In a preferred embodiment the mismatch repair system is the methyl-directed mismatch repair system of *E. coli*.

In a further embodiment, the invention features a method of genomic mismatch scanning such that duplex DNA molecules are subject to exonuclease III digestion only after ligation into monomer circles.

By "ligation into monomer circles" is meant ligation of molecules under conditions of dilute concentration such that ends of the same molecule become ligated. Such a procedure is known to those skilled in the art. In these methods it is advantageous sometimes to separate molecules having mismatches from those which do not. By use of appropriate separation procedures both such populations of molecules can be selected.

Methods Applying Mismatch Repair Stems to Populations of Amplified Molecules

In another aspect, the invention features a method for correcting base pair mismatches in a population of DNA duplexes that have been produced by enzymatic amplification potentially containing one or more base pair mismatches. The method includes contacting the population of DNA duplexes with a DNA methylase and a mismatch repair system such that base pair mismatches are corrected.

By "enzymatic amplification" is meant a reaction by which DNA molecules are amplified. Examples of such reactions include the polymerase chain reaction and reactions utilizing reverse transcription and subsequent DNA amplification of one or more expressed RNA sequences.

By "mismatch repair system" is meant a complete system such that base pair mismatches are detected and corrected.

In a preferred embodiment, the mismatch repair system is the methyl-directed mismatch repair system of *E. coli*. Components of the defined system capable of correcting mismatches include MutH, MutL, and MutS proteins, DNA helicase II, single-strand binding protein, DNA polymerase III holoenzyme, exonuclease I, exonuclease VII or RecI, DNA ligase, ATP and four deoxynucleoside triphosphates.

In a further aspect, the invention features a method for removing DNA molecules containing one or more base pair mismatches in a population of molecules that have been produced by enzymatic amplification potentially containing one or more base pair mismatches. The method includes contacting a population of enzymatically amplified molecules with components of a mismatch repair system under

26

conditions such that one or more components of the repair system form a specific complex with a base pair mismatch contained in a DNA duplex and removing DNA duplexes containing the complex from the population of duplex molecules.

By "complex" is meant the result of specific binding of at least one component of mismatch repair system to a base pair mismatch.

In a preferred embodiment, the mismatch repair system is the *E. coli* methyl-directed mismatch repair system, the component of the system is the MutS protein, the MutS protein is affixed to a solid support and removal of the DNA duplex containing the complex is by binding to this support.

Methods of attachment of proteins to solid support systems and use of those systems to perform chromatography so as to remove specific molecules are well known to those skilled in the art.

In another embodiment, the invention features a method for removing DNA molecules containing one or more base pair mismatches in a population of DNA duplexes that have been produced by enzymatic amplification, potentially containing one or more base pair mismatches. The method comprises the steps of contacting the population of DNA duplexes with components of a mismatch repair system under conditions such that an endonucleolytic incision is made on a newly synthesized strand of a DNA duplex molecule containing a base pair mismatch so that such a molecule cannot produce a full-sized product in a subsequent round of enzymatic amplification.

By "endonucleolytic cleavage" is meant cleavage on the unmethylated strand at a hemimethylate of GATC sequence by components of a mismatch repair system.

By "full sized product" is meant a molecule that includes the entire region of interest that is subject to amplification. Molecules that contain endonucleolytic cleavage cannot be amplified in subsequent rounds to produce full sized product and thus will be eliminated from the final amplified product population.

In a preferred embodiment the mismatch repair system is the methyl-directed mismatch repair system of *E. coli* and the components are MutS, MutL, and Mute proteins, and ATP.

Methods to Remove from a Population Molecules Containing a Base Pair Mismatch

In a further embodiment the invention features a method for removing DNA duplex molecules containing base pair mismatches in a population of heteroduplex DNA molecules produced from different sources. The method comprises contacting the population of DNA duplex molecules potentially containing base pair mismatches with some or all components of a mismatch repair system under conditions such that the component or components form a complex with the DNA having a base pair mismatch, and not with a DNA duplex lacking a base pair mismatch, and removing DNA molecules containing the complex or the product of the complex.

By "product of the complex" is meant a DNA duplex that has incorporated biotinylated nucleotides.

By "some or all components of a mismatch repair system" is meant either a complete mismatch repair system such that the complete reaction is carried out or only the proteins of the system which specifically bind to the mismatch.

In preferred embodiments the mismatch repair system is the methyl-directed mismatch repair system of *E. coli*; some or all protein of the mismatch repair system have been affixed to a solid support and removal by adsorption; the complex interacts with other cellular proteins, and removal

5,702,894

27

of the complex occurs through the interaction; and the conditions include the use of biotinylated nucleotides such that the nucleotides are incorporated into duplex molecules that contained a base pair mismatch and such duplexes are removed by binding to avidin.

By "some or all proteins" is meant, for example, *E. coli* proteins MutS, MutL, and MutH.

By "attached to a solid support" is meant a means, such as by fusion with glutathione transferase, by which a protein is attached to a solid support system and still remains functional.

By "adsorption" is meant specific binding to some or all of the proteins of the mismatch repair system affixed to a solid support so that separation from other molecules that do not bind to the solid support affixed proteins occurs.

By "interacts with other cellular proteins" is meant interaction between mismatch repair system protein or between those proteins and other proteins. For example, the interaction of MutS bound to a duplex DNA containing a mismatch with MutL or RecA.

Kits Containing a Mismatch Repair System

In a preferred embodiment, a kit for correcting base pair matches in duplex DNA molecules including one or more of the following components comprising the following purified components: an aliquot of *E. coli* MutH, MutL, and MutS proteins or species variations thereof, an aliquot of DNA helicase II, an aliquot of single-strand DNA binding protein, an aliquot of DNA polymerase III holoenzyme, an aliquot of exonuclease I, an aliquot of Exo VII or RecJ, an aliquot of DNA ligase, an aliquot of ATP, and an aliquot of four deoxynucleoside triphosphates.

A further embodiment of this aspect of this invention includes an assay kit for eliminating DNA molecules containing one or more base pairing mismatches from a population of heterohybrid duplex molecules formed by base pairing of single-stranded DNA molecules obtained from a first and a second source comprising one or more of the following components, an aliquot of proteins of a mismatch repair system, and an aliquot of a recombinase protein.

By "proteins of a mismatch repair system" are meant proteins that modulate the activity of a recombinase protein.

In a preferred embodiment, the proteins of the mismatch correction system are the MutS and MutL proteins of *E. coli*.

Another aspect of the invention features an assay kit for removing DNA molecules containing one or more base pair mismatches comprising an aliquot of one or more proteins of a mismatch repair system that have been affixed to a column support.

In a preferred embodiment, the protein of the mismatch repair system is the MutS protein of *E. coli*.

Another aspect of the invention features a kit for fractionating a heteroduplex DNA population into two pools, one of which was mismatch-free at the beginning of the procedure, the second of which represents duplexes that contained mispaired bases at the beginning of the procedure. This kit is comprised of one or more of the following components: an aliquot of all components of complete mismatch repair system; an aliquot of biotinylated nucleotides; and an aliquot of avidin or an avidin-based support.

In a preferred embodiment, the mismatch repair system is from *E. coli* and consists of products of the mutH, mutL, and mutS genes, DNA helicase II, single-strand DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, exonuclease VII or RecJ exonuclease, DNA ligase, and ATP.

The following Examples are provided for further illustrating various aspects and embodiments of the present invention and are in no way intended to be limiting of the scope.

28

EXAMPLE 1

DNA Mismatch Correction in a Defined System

In order to address the biochemistry of methyl-directed mismatch correction, the reaction has been assayed in vitro using the type of substrate illustrated in FIG. 1. Application of this method to cell-free extracts of *E. coli* (A. L. Lu, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983) confirmed in vivo findings that methyl-directed repair requires the products of four mutator genes, mutH, mutL, mutS and uvrD (also called mutU), and also demonstrated a requirement for the *E. coli* single-strand DNA binding protein (SSB). The dependence of in vitro correction on mutH, mutL and mutS gene products has permitted isolation of these proteins in near homogeneous, biologically active forms. The MutS protein binds to mismatched DNA base pairs; the MutL protein binds to the MutS-heteroduplex complex (M. Grilley, K. M. Welsh, S. -S. Su, P. Modrich, *J. Biol. Chem.* 264, 1000, 1989); and the 25-kD MutH protein possesses a latent endonuclease that incises the unmethylated strand of a hemimethylated d(GATC) site (K. M. Welsh, A. -L. Lu, S. Clark, P. Modrich, *J. Biol. Chem.* 262, 15624, 1987), with activation of this activity depending on interaction of MutS and MutL with a heteroduplex in the presence of ATP (P. Modrich, *J. Biol. Chem.* 264, 6597, 1989). However, these three Mut proteins together with SSB and the DNA helicase II product of the uvrD (mutU) gene (I. D. Hickson, H. M. Arthur, D. Bramhill, P. T. Emmerson, *Mol. Gen. Genet.* 190, 265, 1983) are not sufficient to mediate methyl-directed repair. Below is described identification of the remaining required components and reconstitution of the reaction in a defined system.

Protein and cofactor requirements for mismatch correction. Methyl-directed mismatch correction occurs by an excision repair reaction in which as much as several kilobases of the unmethylated DNA strand is excised and resynthesized (A.-L. Lin, K. Welsh, S. Clark, S. -S. Su, P. Modrich, *Cold Spring Harbor Symp. Quant. Biol.* 49, 589, 1984). DNA polymerase I, an enzyme that functions in a number of DNA repair pathways, does not contribute in a major way to methyl-directed correction since extracts from a polA deletion strain exhibit normal levels of activity. However extracts derived from a dnaZ⁺ strain are temperature sensitive for methyl-directed repair in vitro (Table 1).

TABLE 1

Requirement for t and g Subunits of DNA Polymerase III Holoenzyme in Mismatch Repair				
Extract genotype	DNA Pol III addition (ng)	Mismatch Correction Activity (fmol/h/mg)		ratio (42°/34°)
		Extract preincubation		
		42°	34°	
dnaZ ⁻	—	8	910.09	
	57 ng	75	1600.47	
dnaZ ⁺	—	150	1600.94	
	57 ng	160	1601.0	

Extracts from strains AX727 (lac thi str⁺ dnaZ20-16) and AX729 (as AX727 except purE dnaZ₂) were prepared as described (A.-L. Lin, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983). Samples (110 µg of protein) were mixed with 0.8 µl of 1M KCl and water to yield a volume of 7.2 µl, and preincubated at 42° or 34° C. for 2.5 minutes. All heated samples were then placed at 34° C. and supplemented with 2.2 µl of a solution containing 0.1 µg (24 fmol)

5,702,894

29

of hemimethylated G-T heteroduplex DNA, 16 ng of MutL protein, 50 ng of MutS protein, and buffer and nucleotide components of the mismatch correction assay (A.-L. Lu, S. Clark, P. Modrich, *Proc. Natl. Acad. Sci. USA* 80, 4639, 1983), DNA polymerase III holoenzyme (57 ng in 0.6 μ l) or enzyme buffer was then added, and incubation at 34° C. was continued for 60 min. Heated extracts were supplemented with purified MutL and MutS proteins because these components are labile at 42° C. Activity measurements reflect the correction of heteroduplex sites.

The *dnaZ* gene encodes the τ and γ subunits of DNA polymerase III holoenzyme (M. Kodaira, S. B. Biswas, A. Kornberg, *Mol. Gen. Genet.* 192, 80, 1983; D. A. Mullin, C. L. Woldringh, J. M. Henson, J. R. Walker, *Mol. Gen. Genet.* 192, 73 1983), and mismatch correction activity is largely restored to heated extracts of the temperature-sensitive mutant strain by addition of purified polymerase III holoenzyme. Since DNA polymerase III holoenzyme is highly processive, incorporating thousands of nucleotides per DNA binding event, the involvement of this activity is consistent with the large repair tracts associated with the methyl-directed reaction.

Additional data indicate that purified MutH, MutL, and MutS proteins, DNA helicase II, SSB, and DNA polymerase III holoenzyme support methyl-directed mismatch correction, but this reaction is inhibited by DNA ligase, an enzyme that is shown below to be required to restore covalent continuity to the repaired strand. This observation led to isolation of a 55-kD stimulatory protein that obviates ligase inhibition. The molecular weight and N-terminal sequence of this protein indicated identity to exonuclease I (G. J. Phillips and S. R. Kushner, *J. Biol. Chem.* 262, 455, 1987), and homogeneous exonuclease I readily substitutes for the 55-kD stimulatory activity (Table 2). Thus, exonuclease I and the six activities mentioned above mediate efficient methyl-directed mismatch correction in the presence of ligase to yield product molecules in which both DNA strands are covalently continuous.

TABLE 2

Stimulation of in vitro Methyl-Directed Correction by Exonuclease I	
Protein added	Mismatch correction (fmol/20 min)
None	1
55-kD protein	18
Exonuclease I	18

Reactions (10 μ l) contained 0.05M HEPES (potassium salt, pH 8.0), 0.02M KCl, 6 mM MgCl₂, bovine serum albumin (0.05 mg/ml), 1 mM dithiothreitol, 2 mM ATP, 100 μ M (each) dATP, dCTP, dGTP, and dTTP, 25 μ M β -NAD⁺, 0.1 μ g of hemimethylated, covalently closed G-T heteroduplex DNA (FIG. 1, methylation on c strand, 24 fmol), 0.26 ng of MutH (K. M. Welsh, A.-L. Lin, S. Clark, P. Modrich, *J. Biol. Chem.* 262, 15624, 1987), 17 ng of MutL (M. Grilley, K. R. Welsh, S.-S. Su, P. Modrich, *J. Biol. Chem.* 264, 1000, 1989), 35 ng of MutS (S.-S. Sin and P. Modrich, *Proc. Natl. Acad. Sci. USA* 83 5057, 1986), 200 ng of SSB (T. R. Lohman, J. R. Green, R. S. Beyer, *Biochemistry* 25, 21, 1986; U.S. Biochemical Corp.), 10 ng of DNA helicase II (K. Kumura and M. Sekiguchi, *J. Biol. Chem.* 259, 1560, 1984), 20 mg of *E. coli* DNA ligase (U.S. Biochemical Corp.), 95 ng of DNA polymerase III holoenzyme (C. McHenry and A. Kornberg, *J. Biol. Chem.* 252, 6478, 1977), and 1 ng of 55-kD protein or exonuclease I (U.S. Biochemi-

30

cal Corp.) as indicated. Reactions were incubated at 37° C. for 20 minutes, quenched at 55° C. for 10 minutes, chilled on ice, and then digested with Xho I or Hind III endonuclease to monitor correction. Repair of the G-T mismatch yielded a only the G-C containing, Xho I-sensitive product.

The requirements for repair of a covalently closed G-T heteroduplex (FIG. 1) are summarized in Table 3 (Closed circular). No detectable repair was observed in the absence of MutH, MutL, or MutS proteins or in the absence of DNA polymerase III holoenzyme, and omission of SSB or exonuclease I reduced activity by 85 to 90 percent.

TABLE 3

Protein and Cofactor Requirements for Mismatch Correction in a Defined System		
Reaction conditions	Mismatch correction (fmol/20 min)	
	Closed Circular Heteroduplex	Open Circular Heteroduplex
Complete	15	17 (No MutH, No ligase)
minus MutH	<1	—
minus MutL	<1	<1
minus MutS	<1	<1
minus DNA polymerase III holoenzyme	<1	<1
minus SSB	2	1.4
minus exonuclease I	2	<1
minus DNA helicase II	16	15
minus helicase II, plus immune serum	<1	<1
minus helicase II, plus pre-immune serum	14	NT
minus Ligase/NAD ⁺	14	NT
minus MgCl ₂	<1	NT
minus ATP	<1	NT
minus dNTP's	<1	NT

Reactions utilizing covalently closed G-T heteroduplex (modification on c strand) were performed as described in the legend to Table 2 except that 1.8 ng of exonuclease I was used. Repair of open circular DNA was performed in a similar manner except that MutH, DNA ligase, and β -NAD⁺ were omitted from all reactions, and the hemimethylated G-T heteroduplex (modification on c strand) had been incised with MutH protein as described in the legend to FIG.

4. When present, rabbit antiserum to helicase II or pre-immune serum (5 μ g protein) was incubated at 0° C. for 20 minutes with reaction mixtures lacking MgCl₂; the cofactor was then added and the assay was performed as above. Although not shown, antiserum inhibition was reversed by the subsequent addition of more helicase II. With the exception of the DNA polymerase III preparation, which contained about 15% by weight DNA helicase II (text) the purity of individual protein fractions was \geq 95%. NT—not tested.

These findings are in accord with previous conclusions concerning requirements of the methyl-directed reaction. However, in contrast to observations in vivo and in crude extracts indicating a requirement for the *uvrD* product, the reconstituted reaction proceeded readily in the absence of the added DNA helicase II (Table 2). Nevertheless, the reaction was abolished by antiserum to homogeneous helicase II, suggesting a requirement for this activity and that it might be present as a contaminant in one of the other proteins. Analysis of these preparations for their ability to restore mismatch repair to an extract derived from a *uvrD* (*mutU*) mutant and for the physical presence of helicase II by immunoblot assay revealed that the DNA polymerase III

5,702,894

31

holoenzyme preparation contained sufficient helicase II (13 to 15 per cent of total protein by weight) to account for the levels of mismatch correction observed in the defined system. Similar results were obtained with holoenzyme preparations obtained from two other laboratories. The purified system therefore requires all the proteins that have been previously implicated in methyl-directed repair.

The rate of correction of the closed circular heteroduplex was unaffected by omission of DNA ligase (Table 3), but the presence of this activity results in production of a covalently closed product. Incubation of a hemimethylated, supercoiled G-T heteroduplex with all seven proteins required for correction in the presence of DNA ligase resulted in extensive formation of covalently closed, relaxed, circular molecules. Production of the relaxed DNA was dependent on MutS (FIG. 2A) and MutL proteins, and the generation of this species was associated with heteroduplex repair (FIG. 2B). Correction also occurred in the absence of ligase, but in this case repair products were open circular molecules, the formation of which depended on the presence of MutS (FIGS. 2A and 2B).

Since MutS has no known endonuclease activity but does recognize mispairs, it is inferred that open circular molecules are the immediate product of a mismatch-provoked excision repair process. Ligase closure of the strand break(s) present in this species would yield the covalently closed, relaxed circular product observed with the complete system.

The set of purified activities identified here as being important in methyl-directed repair support efficient correction. In the experiments summarized in Table 3, the individual proteins were used at the concentrations estimated to be present in the standard crude extract assay for correction as calculated from known specific activity determinations. Under such conditions the rate and extent of mismatch repair in the purified system are essentially identical to those observed in cell-free extracts.

DNA Sites Involved in Repair by the Purified System.

The single d(GATC) sequence within the G-T heteroduplex shown in FIG. 1 is located 1024 base pairs from the mispair. Despite the distance separating these two sites, correction of the mismatch by the purified system responded to the state of modification of the d(GATC) sequence as well as its presence within the heteroduplex (FIG. 3). A substrate bearing d(GATC) methylation on both DNA strands did not support mismatch repair nor did a related heteroduplex in which the d(GATC) sequence was replaced by d(GATT). However, each of the two hemimethylated heteroduplexes were subject to strand-specific correction, with repair in each case being restricted to the unmodified DNA strand. With a heteroduplex in which neither strand was methylated, some molecules were corrected on one strand, and some were corrected on the other. As can be seen, the hemimethylated heteroduplex bearing methylation on the complementary DNA strand was a better substrate than the alternative configuration in which modification was on the viral strand, with a similar preference for repair of the viral strand being evident with the substrate that was unmethylated on either strand. This set of responses of the purified system to the presence and state of modification of d(GATC) sites reproduce effects previously documented in vivo and in crude extract experiments (R. S. Lahue, S. -S. Su, P. Modrich, *Proc. Natl. Acad. Sci. USA* 84, 1482, 1987).

32

TABLE 4

Correction Efficiencies for Different Mismatches.					
Heteroduplex	Markers	C ⁺ V ⁻		C ⁻ V ⁺	
		Rate	Bias	Rate	Bias
C 5'-CTCGA G AGCTT	Xho I	1.2	>18	0.38	>5
V 3'-GAGCT T TCGAA	Hind III				
C 5'-CTCGA G AGCTG	Xho I	1.1	>17	0.38	>6
V 3'-GAGCT G TCGAC	Pvu II				
C 5'-ATCGA T AGCTT	Cla I	1.0	>16	0.24	3
V 3'-TAGCT T TCGAA	Hind III				
C 5'-ATCGA A AGCTT	Hind III	0.88	>20	0.20	>7
V 3'-TAGCT A TCGAA	Cla I				
C 5'-CTCGA A AGCTT	Hind III	0.61	17	0.28	>5
V 3'-GAGCT C TCGAA	Xho I				
C 5'-GTCGA C AGCTT	Sal I	0.60	12	0.23	>4
V 3'-CAGCT T TCGAA	Hind III				
C 5'-GTCGA A AGCTT	Hind III	0.44	>13	0.21	5
V 3'-CAGCT T TCGAA	Sal I				
C 5'-CTCGA C AGCTG	Pvu II	0.04	NS	<0.04	NS
V 3'-GAGCT C TCGAC	Xho I				

Table 4. (Continued) Correction of the eight possible base-base mispairs was tested with the set of covalently closed heteroduplexes described previously including the G-T substrate shown in FIG. 1. With the exception of the mispair and the variations shown at the fifth position on either side, all heteroduplexes were identical in sequence. Each DNA was tested in both hemimethylated configurations under complete reaction conditions (Table 3, closed circular heteroduplex) except that samples were removed at 5-minute intervals over a 20 minute period in order to obtain initial rates (fmol/min). c and v refer to complementary and viral DNA strands, and Bias indicates the relative efficiency of mismatch repair occurring on the two DNA strands (ratio of unmethylated to methylated) as determined 60 minutes after the reaction was started. NS—not significant. With the exception of the C—C heteroduplexes, repair in the absence of MutS protein was less than 20% (in most cases <10%) of that observed in its presence (not shown).

The efficiency of repair by the methyl-directed pathway depends not only on the nature of the mispair, but also on the sequence environment in which the mismatch is embedded (P. Modrich, *Ann. Rev. Biochem.* 56, 435, 1987). To assess the mismatch specificity of the purified system under conditions where sequence effects are minimized, a set of heteroduplexes were used in which the location and immediate sequence environment of each mispair are essentially identical (S. -S. Su, R. S. Lahue, K. G. Au, P. Modrich, *J. Biol. Chem.* 263 6829, 1988). This analysis (Table 4) showed that the purified system is able to recognize and repair in a methyl-directed manner seven of the eight possible base-base mismatches, with C—C being the only mispair that was not subject to significant correction. Table 3 also shows that the seven corrected mismatches were not repaired with equal efficiency and that in the case of each heteroduplex, the hemimethylated configuration modified on the complementary DNA strand was a better substrate than the other configuration in which the methyl group was on the viral strand. These findings are in good agreement with patterns of repair observed with this set of heteroduplexes in *E. coli* extracts (Although the patterns of substrate activity observed in extracts and in the purified system are qualitatively identical, the magnitude of variation observed differs for the two systems. Hemimethylated heteroduplexes modified on the complementary DNA strand are better substrates in both systems, but in extracts such molecules are repaired at about twice the rate of molecules methylated on

5,702,894

33

the viral strand. In the purified system these relative rates differ by factors of 2 to 4. A similar effect may also exist with respect to mismatch preference within a given hemimethylated family. Although neither system repairs C—C, the rates of repair of other mismatches vary by a factors of 1.5 to 2 in extracts but by factors of 2 to 3 in the defined system.).

Strand-specific repair directed by a DNA strand break. Early experiments on methyl-directed repair in *E. coli* extracts led to the proposal that the strand-specificity of the reaction resulted from endonucleolytic incision of an unmethylated DNA strand at a d(GATC) sequence. This idea was supported by the finding that purified Muth protein has an associated, but extremely weak d(GATC) endonuclease that is activated in a mismatch-dependent manner in a reaction requiring MutL, MutS, and ATP. The purified system has been used to explore this effect more completely.

The two hemimethylated forms of the G-T heteroduplex shown in FIG. 1 were incised using high concentrations of purified Muth protein to cleave the unmethylated DNA strand at the d(GATC) sequence (>>pGpApTpC). After removal of the protein, these open circular heteroduplexes were tested as substrates for the purified system in the absence of DNA ligase. Both open circular species were corrected in a strand-specific manner and at rates similar to those for the corresponding covalently closed heteroduplexes (FIG. 4). As observed with closed circular heteroduplexes, repair of the Muth-cleaved molecules required MutL, MutS, SSB, DNA polymerase III holoenzyme, and DNA helicase II (FIG. 4 and open circle entries of Table 2), but in contrast to the behavior of the closed circular substrates, repair of the mismatch within the open circular molecules occurred readily in the absence of Muth protein. Thus prior incision of the unmethylated strand of a d(GATC) site can bypass the requirement for Muth protein in strand-specific mismatch correction.

The nature of the Muth-independent repair was examined further to assess the effect of ligase on the reaction and to determine whether a strand break at a sequence other than d(GATC) can direct correction in the absence of Muth protein (FIG. 5). As mentioned above, a covalently closed G-T heteroduplex that lacks a d(GATC) sequence is not subject to repair by the purified system in the presence (FIG. 3) or absence of DNA ligase. However, the presence of one strand-specific, site-specific break is sufficient to render this heteroduplex a substrate for the purified system in the absence of ligase and Muth protein (FIG. 5). Repair of this open circular heteroduplex was limited to the incised, complementary DNA strand, required presence of MutL and MutS proteins, DNA polymerase III, and SSB, and correction of the molecule was as efficient as that observed with the hemimethylated heteroduplex that had been cleaved by Muth at the d(GATC) sequence within the complementary strand. Although the presence of a strand break is sufficient to permit strand-specific correction of a heteroduplex in the absence of Muth and ligase, the presence of the latter activity inhibited repair not only on the heteroduplex lacking a d(GATC) sequence but also on both hemimethylated molecules that had been previously incised with Muth protein (FIG. 5). This inhibition by ligase was circumvented by the presence of Muth protein, but only if the substrate contained a d(GATC) sequence, with this effect being demonstrable when both types of heteroduplex were present in the same reaction (FIG. 5, last column). This finding proves that Muth protein recognizes d(GATC) sites and is consistent with the view that the function of this protein in mismatch correction is the incision of the unmethylated strand at this sequence.

34

EXAMPLE 2

Purification of MutY Protein

Purification of MutY Protein *E. coli* RK1517 was grown at 37° C. in 170 liters of L broth containing 2.5 mM KH_2PO_4 , 7.5 mM Na_2HPO_4 (culture pH=7.4) and 1% glucose. The culture was grown to an A_{500} of 4, chilled to 10° C. and cells were harvested by continuous flow centrifugation. Cell paste was stored at 70° C. A summary of the MutY purification is presented in Table 1. Fractionation procedures were performed at 0°–4° C., centrifugation was at 13,000×g, and glycerol concentrations are expressed as volume percent.

Frozen cell paste (290 g) was thawed at 4° C., resuspended in 900 ml of 0.05M Tris-HCl (pH 7.5), 0.1M NaCl, 1 mM dithiothreitol, 0.1 mM EDTA, and cells were disrupted by sonication. After clarification by centrifugation for 1 hr, the lysate (Fraction I, 970 ml) was treated with 185 ml of 25% streptomycin sulfate (wt/vol in 0.05M Tris-HCl (pH 7.5), 0.1M NaCl, 1 mM dithiothreitol, 0.1 mM EDTA) which was added slowly with stirring. After 30 min of additional stirring, the solution was centrifuged for 1 h, and the supernatant (1120 ml) was treated with 252 g of solid ammonium sulfate which was added slowly with stirring. After 30 min. of additional stirring, the precipitate was collected by centrifugation for 1 h, resuspended to a final volume of 41 ml in 0.02M potassium phosphate (pH 7.5), 0.1 mM EDTA, 10% (vol/vol) glycerol, 1 mM dithiothreitol, and dialyzed against two 2 l portions of 0.02M potassium phosphate (pH 7.5), 0.1M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol (2 h per change). The dialyzed material was clarified by centrifugation for 10 min to yield Fraction II (45 ml).

Fraction II was diluted 10-fold into 0.02M potassium phosphate (pH 7.5), 0.1M EDTA, 1 mM dithiothreitol, 10% glycerol so that the conductivity of the diluted solution was comparable to that of the dilution buffer containing 0.1M KCl. The solution was performed on small aliquots of Fraction II, and diluted samples were immediately loaded at 1 ml/min onto a 14.7 cm×12.6 cm² phosphocellulose column equilibrated with 0.02 M potassium phosphate (pH 7.5), 0.1M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. The column was washed with 400 ml of equilibration buffer, and developed with a 2 liter linear gradient of KCl (0.1 to 1.0M) in 0.02M potassium phosphate (pH 7.5), 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. Fractions containing MutY activity, which eluted at about 0.4M KCl, were pooled (Fraction III, 169 ml).

Fraction III was dialyzed against two 500 ml portions of 5 mM potassium phosphate (pH 7.5), 0.05M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol (2 h per change) until the conductivity was comparable to that of the dialysis buffer. After clarification by centrifugation at for 10 min, the solution was loaded at 0.5 ml/min onto a 21 cm×2.84 cm² hydroxylapatite column equilibrated with 5 mM potassium phosphate, pH 7.5, 0.05M KCl, 0.1 mM EDTA, 1 mM dithiothreitol, 10% glycerol. After washing with 130 ml of equilibration buffer, the column was eluted with a 600 ml linear gradient of potassium phosphate (5 mM to 0.4M, pH 7.5) containing 0.05M KCl, 1 mM dithiothreitol, 10% glycerol. Fractions eluting from the column were supplemented with EDTA to 0.1 mM. Peak fractions containing 60% of the total recovered activity, which eluted at about 0.1M potassium phosphate, were pooled (Fraction IV, 24 ml). The remaining side fractions contained impurities which could not be resolved from MutY by MonoS chromatography.

5,702,894

35

Fraction IV was diluted by addition of an equal volume of 0.1 mM EDTA, 1 dithiothreitol, 10% glycerol. After clarification by centrifugation for 15 min, diluted Fraction IV was loaded at 0.75 ml/min onto a Pharmacia HR 5/5 MonoS FPLC column that was equilibrated with 0.05M sodium phosphate (pH 7.5), 0.1M NaCl, 0.1 mM EDTA, 0.5 mM dithiothreitol, 10% glycerol. The column was washed at 0.5 ml/min with 17 ml of equilibration buffer and developed at 0.5 ml/min with a Ex 2/ Table 1

TABLE 1

Purification of MutY protein from 290 g of <i>E. coli</i> RK1517		Total Protein mg	Specific Activity units/mg	Yield Percent
Fraction	Step			
I	Extract	10,900	40	(100)
II	Ammonium sulfate	1,350	272	84
III	Phosphocellulose	66	10,800	160
IV	Hydroxylapatite	1.4	136,000	44
V	MonoS	0.16	480,000	18

Specific A*G to C-G mismatch correction in cell-free extracts was determined as described previously (Au et al. 1988), except that ATP and glutathione were omitted from the reaction and incubation was for 30 min instead of 1 h. For complementation assays, each 0.01 ml reaction contained RK1517-Y33 extract (mutS mutY) at a concentration of 10 mg/ml protein. One unit of MutY activity is defined as the amount required to convert 1 fmol of A*G mismatch to C-G base pair per h under complementation conditions.

20 ml linear gradient of NaCl (0.1 to 0.4M) in 0.05M sodium phosphate (pH 7.5), 0.1 mM EDTA, 0.5 mM dithiothreitol, 10% glycerol. Fractions with MutY activity, which eluted at approximately 0.2M NaCl, were pooled (Fraction V, 2.6 ml). Fraction V was divided into small aliquots and stored at -70° C.

Assay for MutY-dependent, A*G-specific Glycosylase

DNA restriction fragments were labeled at either the 3' or 5' ends with ³²P. Glycosylase activity was then determined in 0.01 ml reactions containing 10 ng end-labeled DNA fragments, 0.02M Tris-HCl, pH 7.6, 1 mM EDTA, 0.05 mg/ml bovine serum albumin, and 2.7 ng MutY. After incubation at 37° C. for 30 min, the reaction mixture was treated with 2.5x10⁻³ units of HeLa AP endonuclease II in the presence of 11 mM MgCl₂ and 0.005% Triton X-100 for 10 min at 37° C. Reactions were quenched by the addition of an equal volume of 80% formamide, 0.025% xylene cyanol, 0.025% bromophenol blue, heated to 80° C. for 2 min, and the products analyzed on an 8% sequencing gel. Control reactions contained either no MutY, no A*G mismatch or no AP endonuclease II.

Strand cleavage at the AP site generated by MutY could also be accomplished by treatment with piperidine instead of treatment with AP endonuclease II. After incubation for 30 min. at 37° C. with MutY as described above, the reaction mixture was precipitated with ethanol in the presence of carrier tRNA, then resuspended in 1M piperidine and heated at 90° C. for 30 min. After two additional ethanol precipitations, changing tubes each time, the pellet was resuspended in a minimum volume of water to which was added an equal volume of 80% formamide, 0.025% xylene cyanol, 0.025% bromophenol blue. The products were then analyzed on an 8% sequencing gel.

EXAMPLE 3

Genetic Mapping Point Mutations in the Human Genome

The full novelty and utility of the present invention may be further appreciated by reference to the following brief

36

description of selected specific embodiments which advantageously employ various preferred forms of the invention as applied to a common problem in genetic mapping of point mutations in the human genome. In the course of constructing gene linkage maps, for example, it is frequently desirable to compare the sequence of a cloned DNA fragment with homologous sequences in DNA extracted from a human tissue sample. Substantially all base pairs in the entire homologous sequence of the cloned DNA fragment are compared to those of the human tissue DNA, most advantageously in a single test according to the present invention, merely by contacting both strands of the human tissue DNA molecule with both radiolabeled complementary strands of the second DNA molecule under conditions such that base pairing occurs, contacting the resulting DNA duplexes with the *E. coli* MutS protein that recognizes substantially all base pair mismatches under conditions such that the protein forms specific complexes with its cognate mispairs, and detecting the resulting DNA:protein complexes by contacting the complexes with a membranous nitrocellulose filter under conditions such that protein:DNA complexes are retained while DNA not complexed with protein is not retained, and measuring the amount of DNA in the retained complexes by a standard radiological methods or by utilizing any of the other methods of the invention; e.g., altered electrophoretic mobility, or detection by use of antibodies.

If the above detection test indicates the presence of sequence differences between the human tissue DNA and the cloned DNA and localization is required, or, in the alternative, if such differences are suspected and localization as well as detection of them is desired in a first analysis, the another method of this invention may be applied for these purposes. An embodiment of this aspect of the invention that may be most advantageously employed comprises the steps of contacting both strands of the human tissue DNA molecule with both radiolabeled complementary strands of the second DNA molecule (usually without separation from the cloning vector DNA) under conditions such that base pairing occurs, contacting the resulting DNA duplexes with MutHLS to produce a GATC cleavage reaction or a modified form of MutS protein of *E. coli* to which is attached an hydroxyl radical cleaving function under conditions such that the radical cleaving function cleaves both strands of the DNA within about 20 base pairs of substantially all DNA base mispairs. In the absence of any DNA base mispairs in the DNA duplexes comprising complementary strands of the human tissue and cloned DNAs, no DNA fragments smaller than the cloned DNA (plus vector DNA, if still attached) would be detected. Determination of the location of any double-stranded DNA cleavages by the modified MutS protein to within a few kbp or less of some restriction enzyme cleavage site within the cloned DNA is determined by standard restriction enzyme mapping approaches. If greater precision in localization and identification of a single base difference is desired, sequencing could be confined to those particular fragments of cloned DNA that span at least one base sequence difference localized by this method and are cleaved by a restriction enzyme at the most convenient distance of those sequence differences for direct sequencing.

The examples herein can be changed to make use of other methods of separation to identify mismatches, such as a filter-binding assay, as well as the nicking reaction with MutS and MutL. While large (at least 20 kbp) or small DNA molecules can be used in these methods those of between 1-10 kbp are preferred.

EXAMPLE 4

DNA Mismatch Detection Kit

Kit contains MutS protein, dilution buffer, annealing buffer, reagents to generate complementary and mismatched

5,702,894

37

control duplexes and filter binding protocol. It can be used to detect single-base mismatches in oligonucleotides.

MutS kit components:

MutS protein in storage buffer: 50 mM HEPES pH7.2, 100 mM KCl, 1 mM EDTA, 1 mM DTT;

MutS1: 16 mer oligonucleotide GATCCGTCGACCT-GCA (all such oligonucleotides are written 5' to 3' herein) in water (2 μ M);

MutS2: 16mer oligonucleotide TGCAGGTCGACG-GATC 1 μ M in annealing buffer 1 μ M: 20 mM Tris/HCl pH 7.6, 5 mM, MgCl₂, 0.1 mM DTT, 0.01 mM EDTA;

MutS3: 16 mer oligonucleotide TGCAGGTTGACG-GATC 1 μ M in annealing buffer;

Assay buffer/annealing buffer/wash buffer, 20 mM Tris/HCl pH 7.6, 5 mM MgCl₂, 0.1 mM DTT, 0.01 mM EDTA;

Protein storage/dilution buffer: 50 mM HEPES pH 7.2, 100 mM KCl, 1 mM EDTA, 1mM DTT.

The DNA mismatch detection kit contains three 16-mer oligonucleotides labeled MUTS1, MUTS2, and MUTS3 for testing the performance of MutS protein. When MUTS1 and MUTS2 are annealed, a perfectly matched duplex results. When MUTS1 and MUTS3 are annealed, a duplex containing a single G-T mismatch results. These serve as control substrates for MutS binding.

Kinase Labeling of MUTS1 Oligonucleotide

This protocol uses half the amount of oligonucleotide contained in the kit. To a microcentrifuge tube on ice add the following:

MUTS1 Oligonucleotide (2 μ M)	15 μ l (30 pmoles)
10X T4 Polynucleotide Kinase Buffer	3 μ l
³² P-ATP (3000 Ci/nmole)	1 μ l
ATP (10 μ M)	2.5 μ l
Sterile dH ₂ O	7.5 μ l
T4 Polynucleotide Kinase (30 units/ μ l)	1 μ l (30 units)

Incubate the reaction mixture for 10 min at 37° C. Then incubate 10 min at 70° C. Spot two independent 1 μ l aliquots of the mixture on a SureCheck TLC plate and also spot a dilution of ³²P-ATP (1:30 in water) in a separate lane and run with the elution mixture. Expose the developed plate to X-ray film for 5 min. Scrape all radioactive spots from both experimental lanes of the plate and count them in a liquid scintillation counter to determine the % incorporation of label. This value is typically 40–60%. If a significant labeled ATP spot is present in the kinase reaction lanes on the plate, the labeled oligonucleotide must be purified before use (TLC or gel), since ³²P-ATP will contribute to background in the filter binding assay. In our experience, this is usually not necessary.

Keep in mind that the MUTS1 oligo stock is 2 pmol/ μ l and that the final concentration should be 1 pmol/ μ l. It is critical that this final concentration be as exact as possible, since the concentration determines the amount of MUTS1 in the next (annealing) step and hence, the amount of DNA available for binding by the protein.

Annealing Reactions

Two separate reactions are carried out: MUTS1/MUTS2 and MUTS1/MUTS3. In both cases, the ³²P-labeled MUTS1 from Step 1 is used.

38

Complementary		Mismatched	
MUTS1 (kinased)	14 μ l = 14 pmoles	MUTS1 (kinased)	14 μ l = 14 pmoles
MUTS2 (1 μ M)	28 μ l = 28 pmoles	MUTS3 (1 μ M)	28 μ l = 28 pmoles
annealing buffer	28 μ l	annealing buffer	28 μ l
	70 μ l		70 μ l

1. Heat each mixture for 10 min at 70° C.
2. Incubate for 30 min at room temperature.
3. Hold on ice until ready to use.

The molar ratio of MUTS2/MUTS1 and MUTS3/MUTS1 is 2:1 in the above reactions and this should be maintained for optimal results. Lowering the ratio of unlabeled to labeled strand may lead to very high background in the filter binding assay, presumably caused by sticking of labeled ssDNA to nitrocellulose.

Assay of MutS Binding by the Gel Shift Method

The binding of MutS to mismatches can be assessed using the technique of Gel Shift Mobility Assay (GSMA), a useful tool to identify protein-DNA interactions which may regulate gene expression. Below is a protocol for performing GSMA on the MUTS1/MUTS3 mismatched duplex contained in the mismatch detection kit. Optimum conditions may vary depending on the particular mismatch being detected or the length of the oligonucleotide.

All binding reactions should be carried out on ice. The total binding reaction volume is 10 μ l. Add 4 μ l of a MutS protein dilution (prepared using dilution buffer in the kit) containing 0.5–5 pmoles (0.125–1.25 units) of MutS protein (1 pmol=97 ng) to 6 μ l = 1.2 pmoles of ³²P-labeled MUTS1/MUTS3 heteroduplex. Also add comparable amounts of MutS protein to labeled MUTS1/MUTS2 matched duplex to serve as a control. A control incubation consisting only of mismatched heteroduplex (no MutS protein) should also be run. Incubate all reactions on ice for 30 min.

To 3 μ l of the DNA/MutS mixture from each incubation add 1 μ l of a 50% w/v sucrose solution.

Load 2 μ l of the mixture from Step 2 onto a 6% non-denaturing polyacrylamide gel prepared in Tris-acetate-EDTA (TAE) buffer (Sambrook et al., "Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor Laboratory, New York (1989)) to which MgCl₂ has been added to a final concentration of 1 mM and run the gel at 10 V/cm and 4° C. in TAE buffer containing 1 mM MgCl₂ until bromophenol blue dye (loaded into an adjacent well) has migrated approximately half way down the gel. The presence of Mg++ in the gel and running buffer is critical for optimal results in the GSMA assay of MutS protein.

Filter Binding Assay

The total binding reaction volume is 10 μ l. It consists of 6 μ l, or 1.2 pmoles, of duplex DNA and 4 μ l of a MutS protein dilution containing 0.5–5 pmoles (0.125–1.25 units) of MutS protein (1 pmol=97 ng). Each type of duplex, complementary and mismatched, should be assayed in duplicate or triplicate along with a no protein control for each annealing, which will serve as the background to subtract.

In order to use the filter binding assay it will be necessary to make up additional annealing buffer for use in the washing step. Add 20 ml of 1M Tris-HCl, pH 7.6, 5 ml of 1M MgCl₂, 0.1 ml of 1M DTT, and 0.02 ml of 0.5M EDTA to distilled water and bring the volume to 1 liter.

For each binding assay, add the following to a 0.5 ml microcentrifuge tube on ice:

MUTS1/MUTS2 (Control) OR

5,702,894

39

MUTS1/MUTS3 (Mismatched)Annealing Mixture 6 μ l

Set up the filtration apparatus and presoak the nitrocellulose filters in annealing buffer.

Add 4 μ l of MutS protein dilution to the annealing mixtures on ice. Also include no protein controls for each annealing.

After 30 minutes, begin filtration of samples. Caution, use a slow rate of filtration. It should take at least a second or two for the 10 μ l sample to filter.

Immediately wash the filters with 5 ml each of cold annealing buffer. This should take 20–30 seconds.

Place the filters in liquid scintillation vials, add fluid and count for 2 minutes each.

Determine the input cpm for each annealing as follows: To 6 μ l of annealing mixture, add 54 μ l of water and count 2–3 aliquots of 6 μ l each in scintillation fluid. The input cpm is then 10X the average of the cpm of the dilution.

Determine the cpm/pmol of DNA as follows:

$$\frac{\text{cpm of 6 } \mu\text{l aliquot} \times \text{dilution} \times \text{fraction of label incorporate}}{\text{pmol of DNA in annealing reaction}}$$

A 6 μ l annealing contains 1.2 pmoles of DNA

A typical kinase reaction may give 42% incorporation (determined previously)

A 6 μ l aliquot of 10X dilution may be 10,600 cpm

$$\frac{10,600 \times 10 \times 0.42}{1.2} = 37,100 \text{ cpm/pmol DNA}$$

Determine the pmoles of DNA bound by various pmoles of MutS. First, determine the pmoles of MutS protein in a binding reaction.:

$$\frac{\text{concentration of MutS} \times \text{volume of protein added}}{\text{molecular weight of MutS} \times \text{dilution factor}}$$

Example: If 4 μ l of a 6X dilution of MutS at 250 μ g/ml is used, then:

$$\frac{250 \text{ ng}/\mu\text{l} \times 4 \mu\text{l}}{97 \text{ ng/pmol} \times 6} = 1.72 \text{ pmoles of MutS in reaction}$$

Then, determine the pmoles of DNA bound:

$$\frac{\text{cpm retained on filter with MutS protein} - \text{cpm on no protein filter}}{\text{cpm/pmol of DNA}}$$

Example: One gets 15,470 cpm on the filter with MutS and 340 cpm with no protein

$$\frac{15,470 \text{ cpm} - 340 \text{ cpm}}{37,100 \text{ cpm/pmol}} = 0.408 \text{ pmoles of DNA bound}$$

Determine the number of pmoles of MutS required to bind 1 pmole of DNA (i.e., a unit of MutS). In the above example, 1.72 pmoles of MutS bound 0.408 pmoles of DNA, such that one unit = $1.72/0.408 = 4.2$ pmoles MutS per mole DNA.

EXAMPLE 5**Effects of MutS and MutL on RecA-catalyzed Strand Transfer**

A model system used to evaluate MutS and MutL effects on RecA catalyzed strand transfer is depicted in FIG. 6. The assay for RecA-catalyzed strand transfer between homolo-

40

gous and quasi-homologous DNA sequences employed the three strand reaction in which one strand from a linear duplex DNA is transferred to an homologous, single-stranded DNA circle (Cox, 78 *Proc. Natl. Acad. Sci. USA* 343, 1981. These experiments exploited the previous observation that RecA is able to support strand transfer between related fd and M13 DNAs (Bianchi et al., 35 *Cell* 511, 1983; DasGupta et al., 79 *Proc. Natl. Acad. Sci. USA* 762, 1982, which are approximately 97% homologous at the nucleotide level. The vast majority of this variation is due to single base pair changes.

Results of experiments on the effects of MutS and MutL on RecA-catalyzed strand transfer between homologous and quasi-homologous DNA sequences are shown in FIG. 7. Reactions (50 μ l) contained 50 mM HEPES (pH 7.5), 12 mM MgCl_2 , 2 mM ATP, 0.4 mM dithiothreitol, 6 mM phosphocreatine, 10 U/ml phosphocreatine kinase, 0.6 nM single-stranded circular DNA (molecules), 7.6 μ g RecA protein, 0.54 μ g SSB, and MutS or MutL as indicated. Reactions were allowed to preincubate at 37° C. for 10 minutes, strand exchange was initiated by addition of linear duplex fd DNA (Rf DNA linearized by cleavage with HpaI, 0.6 nM final concentration as molecules), and incubation continued for 70 minutes. MutS or MutL was added 1 minute prior to addition of duplex DNA. Sample (50 μ l) were quenched by addition of EDTA (25 mM), sodium dodecyl sulphate (0.1%), and proteinase K (150 μ g/ml), followed by incubations at 42° C. for 30 minutes.

The presence of MutS or MutL was without significant effect on strand transfer between linear duplex fd DNA and circular fd single-strands, MutS did inhibit strand transfer between quasi-homologous linear duplex fd DNA and M13 single-strands. Similar results were obtained for strand transfer between duplex M13 DNA and single-stranded fd (data not shown). In contrast, MutL alone did not significantly alter the yield of circular duplex product formed by RecA catalyzed strand transfer between these different DNAs.

EXAMPLE 6**MutL Potentiation of MutS Block to Strand Transfer**

Results of experiments on the MutL potentiation of the MutS block to strand transfer in response to mismatched base pairs are shown in FIG. 8. Reaction mixtures (210 μ l) contained 50 mM PREPES (pH 7.5), 12 mM MgCl_2 , 2 mM ATP, 0.4 mM dithiothreitol, 6 mM phosphocreatine, 10 U/ml phosphocreatine kinase, 0.6 nM (molecules) single-stranded circular DNA, 32 μ g recA protein, and 2.3 μ g SSB. Reactions were preincubated for 10 minutes at 37° C. and strand exchange initiated by addition of duplex fd DNA (Rf DNA linearized by cleavage with HpaI, 0.6 nM final concentration as molecules). When present, MutS (2.9 μ g) and/or MutL (1.3 μ g) were added 1 minute prior to addition of duplex DNA. Samples were removed as indicated times and quenched as described in Example 5.

MutL potentiates the inhibition of heteroduplex formation that is observed with MutS. Formation of full length, circular heteroduplex product is virtually abolished in the presence of MutS and MutL. Heteroduplex formation between perfectly homologous strands occurred readily in the presence of either or both proteins.

EXAMPLE 7**MutS and MutL Block of Branch Migration**

While MutS and MutS along with MutL blocked formation of fully duplex, circular fd-M13 product, some strand

5,702,894

41

transfer did occur in these reactions as demonstrated by the occurrence of strand transfer "intermediates" that migrated more slowly in agarose gels than fully duplex, nicked circular product (data not shown). The nature of these structures was examined using the S1 nuclease procedure of Cox and Lehman to evaluate mean length of stable heteroduplex formation. This analysis is shown in FIG. 9.

Reaction mixtures (510 µl) contained 50 mM HEPES (pH 7.5), 12 mM MgCl₂, 2 mM ATP, 0.4 mM dithiothreitol, 6 mM phosphocreatine, 10 U/mL phosphocreatine kinase, 0.6 nM single-stranded circular DNA (molecules), 77 µg RecA protein, 5.5 µg SSB, and when indicated 6.9 µg MutS and 3.2 µg MutL. Reactions were allowed to preincubate at 37° C. for 10 minutes, strand exchange was initiated by addition of linear duplex [³H]M13 DNA (Rf DNA linearized by cleavage with HpaI, 0.6 nM final concentration as molecules). MutS or MutL was added 1 minute prior to addition of M13 duplex DNA. Samples (100 µl) were taken as indicated, quenched with sodium dodecyl sulphate (0.8%), and extracted with phenol:chloroform:isoamyl alcohol (24:24:1) equilibrated with 10 mM Tris-HCl, pH 8.0, 0.1 mM EDTA. The organic phase was back-extracted with 0.5 volume of 50 mM HEPES, pH 5.5. Aqueous layers were combined washed with H₂O-saturated ether, and relieved of residual ether by 30 minutes incubation at 37° C. The mean length of stable heteroduplex was then determined using S1 nuclease (10 U/ml) according to Cox and Lehman (Cox, 1981 supra).

Although some strand transfer occurs between fd and M13 DNA s in the presence of MutS and MutL, heteroduplex formation is restricted to about one kilobase of the 6.4 kilobase possible. The MutS and MutL effects on recombination are due, at least in part, to their ability to control branch migration reaction in response to occurrence of mismatched base pairs.

Other embodiments are within the following claims.

What is claimed is:

1. A method for removing DNA molecules containing one or more base pair mismatches in a population of DNA duplexes that have been produced by enzymatic amplification, potentially containing one or more base pair mismatches, comprising the steps of:

contacting said population of DNA duplexes with a mismatch repair system under conditions such that one or more components of said mismatch repair system form a specific complex with a base pair mismatch contained in a DNA duplex having a base pair mismatch, and removing said DNA duplex containing said complex or the product of said complex from the population of duplex molecules without the use of an additional agent capable of enzymatic digestion of said DNA duplexes which is not a component of said mismatch repair system, prior to said removal.

42

2. A method for removing DNA duplex molecules containing one or more base pair mismatches in a population of heteroduplex DNA molecules produced from different sources, comprising the steps of:

contacting said population of DNA duplex molecules potentially containing base pair mismatches with some or all components of a mismatch repair system under conditions such that said component or components form a complex with DNA duplex molecules containing a base pair mismatch, and not with DNA duplex molecules lacking a base pair mismatch, and

removing DNA molecules containing said complex or the product of said complex from said population of heteroduplex DNA molecules without the use of an additional agent capable of enzymatic digestion of said DNA duplexes which is not a component of said mismatch repair system prior to said removal.

3. The method of claim 2, wherein some or all proteins of the mismatch repair system have been affixed to a solid support and removal of said complex is by adsorption.

4. The method of claim 2, wherein said complex specially binds with other cellular proteins and removal of said complex occurs through said interaction.

5. Assay kit for eliminating DNA molecules containing one or more base pair mismatches from a population of heterohybrid duplex molecules formed by base pairing of single-stranded DNA molecules obtained from a first and a second source comprising:

an aliquot of proteins of a mismatch repair system,

an aliquot of a recombinase protein whose ability to catalyze duplex formation is inhibited by said proteins of a mismatch repair system, and

a means for removing said DNA molecules containing one or more base pairing mismatches.

6. The kit of claim 5, wherein the proteins of the mismatch correction system are the MutS and MutL proteins of *Escherichia coli*.

7. Kit for fractionating a heteroduplex DNA population into two parts, one consisting of duplexes which contain a mismatch prior to fractionating and the other consisting of duplexes that do not contain a mismatch prior to fractionating comprising:

an aliquot of components of a mismatch repair system;

an aliquot of biotinylated nucleotides; and

an aliquot of avidin or avidin linked to a support.

8. The kit of claim 7, wherein the mismatch repair system is from *Escherichia coli* and consists of products of the mutH, mutL, and mutS genes, DNA helicase II, single-strand DNA binding protein, DNA polymerase III holoenzyme, exonuclease I, Exo VII exonuclease or RecJ exonuclease, DNA ligase, and ATP.

* * * * *

EXHIBIT E



US005750335A

United States Patent [19]
Gifford

[11] **Patent Number:** **5,750,335**
[45] **Date of Patent:** **May 12, 1998**

[54] **SCREENING FOR GENETIC VARIATION**[75] **Inventor:** David K. Gifford, Weston, Mass.[73] **Assignee:** Massachusetts Institute of Technology, Cambridge, Mass.[21] **Appl. No.:** 52,157[22] **Filed:** Apr. 22, 1993**Related U.S. Application Data**

[63] Continuation-in-part of Ser. No. 874,192, Apr. 24, 1992, abandoned.

[51] **Int. Cl.⁶** C12Q 1/68; C12P 19/34; C07K 14/195; C07K 17/00[52] **U.S. Cl.** 435/6; 435/91.2; 435/810; 536/25.4; 530/350; 530/412; 530/810; 935/77; 935/78[58] **Field of Search** 435/6, 5, 91, 810; 536/23.1, 25.4; 935/77, 78; 530/810, 412, 350[56] **References Cited****U.S. PATENT DOCUMENTS**

4,535,058	8/1985	Weinberg et al.	435/6
4,556,643	12/1985	Paau et al.	436/501
4,725,537	2/1988	Fritsch et al.	435/6
4,794,075	12/1988	Ford et al.	435/6
5,045,450	9/1991	Thilly et al.	435/6
5,459,039	10/1995	Modrich	435/6

FOREIGN PATENT DOCUMENTS

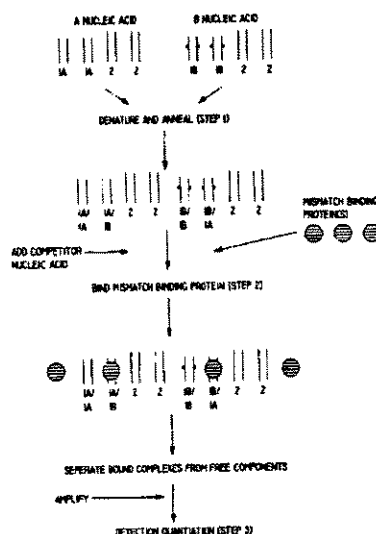
0 265 244 A3	4/1988	European Pat. Off.
0 407 789 A1	1/1991	European Pat. Off.
0 412 883 A1	2/1991	European Pat. Off.
WO 90/13668	1/1991	WIPO
WO 91/00925	1/1991	WIPO
WO 91/13075	9/1991	WIPO
WO 93/02216	2/1993	WIPO

OTHER PUBLICATIONSBlackwell et al., "Differences and Similarities in DNA-Binding Preferences of MyoD and E2A protein Complexes Revealed by Binding Site Selection", *Science*, vol. 250, pp. 1104-1110, 1990.Cotton, "Detection of Single Base Changes in Nucleic Acid", *Advances in Genome Biology*, vol. 1, pp. 253-300, 1991.Hochuli et al., "Genetic Approach to Facilitate Purification of Recombinant Proteins with a Novel Metal Chelate Absorbent", *Bio/Technology*, Nov. 1988, pp. 1321-1325. The FLAG Biosystem, International Biotechnologies, Inc., 1991.McKay, "Binding of a Simian Virus 40 T Antigen-related Protein to DNA", *J. Mol. Biol.* (1981) 145, pp. 471-488.

(List continued on next page.)

Primary Examiner—Carla J. Myers*Attorney, Agent, or Firm*—Testa, Hurwitz & Thibault, LLP[57] **ABSTRACT**

Disclosed is a method of genetic screening for a nucleotide variation, the method including the steps of (A) providing a mixture of nucleic acids comprising heteroduplex nucleic acids and excess homoduplex nucleic acids, wherein each said heteroduplex comprises a test nucleic acid strand isolated from an organism and a reference nucleic acid strand, each said heteroduplex also comprising a mismatched nucleotide pair, wherein said excess homoduplex nucleic acids are generated by reannealing of a first test or reference nucleic acid strand with a fully complementary second test or reference nucleic acid strand; (B) subjecting said mixture to a mismatch binding protein under conditions which promote binding to form a heteroduplex/binding protein complex; and (C) detecting the presence of said mismatched nucleotide pair as an indication of the presence of genetic variation between said test and reference nucleic acids.

70 Claims, 10 Drawing Sheets

5,750,335

Page 2

OTHER PUBLICATIONS

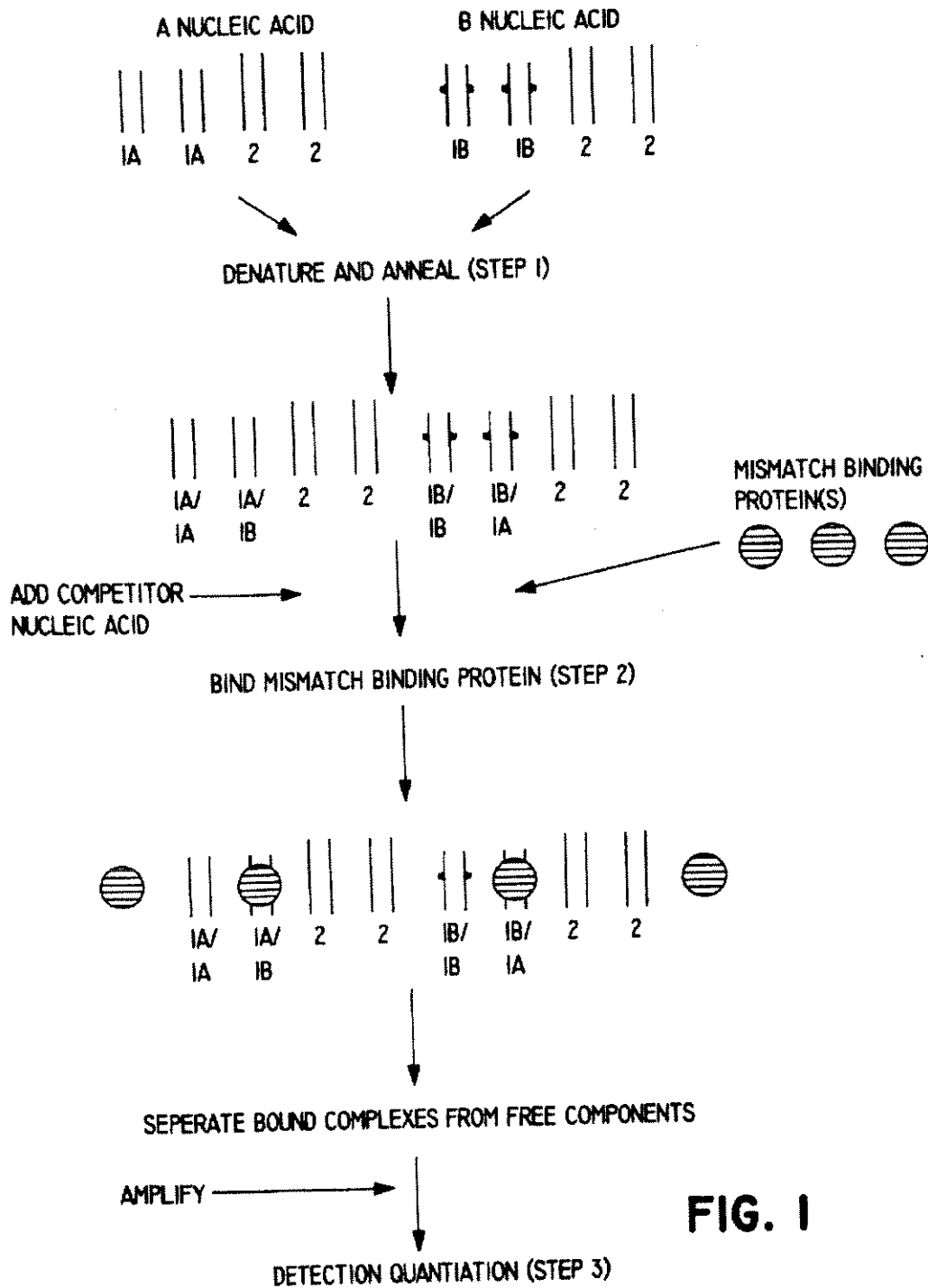
- Mankovich et al., "Nucleotide Sequence of the *Salmonella typhimurium* mutL Gene Required for Mismatch Repair: Homology of MutL to HexB of *Streptococcus pneumoniae* and to PMS1 of the Yeast *Saccharomyces Cerevisiae*", J. Bacteriol., Oct. 1989, vol. 171, No. 10 pp. 5326-5331.
- Protein Fusion and Purification System, New England Biolabs Catalog, 1990-1991, pp. 68-69.
- Potter et al., "A 'Southern Cross' Method for the Analysis of Genome Organization and the Localization of Transcription Units", Gene, 48 (1986) pp. 229-239.
- Yokota et al., "Differential Cloning of Genomic DNA: Cloning of DNA with an Altered Primary Structure by in-gel Competitive Reassociation", Proc. Natl. Acad. Sci. USA, vol. 87, pp. 6398-6402, Aug. 1990.
- Vershon et al., "Isolation and Analysis of Arc Repressor Mutants: Evidence for an Unusual Mechanism of DNA Binding", Proteins: Structure, Function, and Genetics, 1:302-311 (1986).
- Kinzler et al., "Whole Genome PCR: Application to the Identification of Sequences Bound by Gene Regulatory Proteins", vol. 17, No. 10, (1989), pp. 3645-3653.
- Lisitsyn et al., "Cloning the Differences Between Two Complex Genomes", Science, vol. 259, (1993) pp. 946-951.
- Su et al., "*Escherichia coli* mutS-encoded Protein Binds to Mismatched DNA Base Pairs", Proc. Natl. Acad. Sci. USA, vol. 83, pp. 5057-5061, Jul. 1986.
- Harber, "Genetic and Biochemical Analyses of the *Salmonella typhimurium* MutS Mismatch Repair Protein", (1990) pp. 1-96.
- Affinity Chromatography, Pharmacia LKB Biotechnology Catalog, 1992, pp. 72-104.
- Pang et al., "Identification and Characterization of the mutL and mutS Gene Products of *Salmonella typhimurium* LT2", J. Bacteriol., Sep. 1985, pp. 1107-1015, vol. 163, No. 3.
- Cotton, "Detecting of Single Base Changes in Nucleic Acid", JAI Press 1991.
- Stephenson et al., "Selective Binding to DNA Base Pair Mismatches by Proteins from Human Cells", The Journal of Biological Chemistry, (1989) vol. 264, No. 35, pp. 21177-21182.
- Jiricny et al., "A Human 200-kDa Protein Binds Selectively to DNA Fragments Containing GT Mismatches", Proc. Natl. Acad. Sci. USA, vol. 85, pp. 8860-8864, Dec. (1988).
- Modrich, "Mechanisms and Biological Effects of Mismatch Repair", Annu. Rev. Genet. (1991) 25:229-53.
- Jiricny et al., "Mismatch-containing oligonucleotide duplexes bound by the *E. coli* mutS-encoded protein", vol. 16, No. 16 (1988), pp. 7843-7853.
- Roberts et al., "Detection of Novel Genetic Markers by Mismatch Analysis", vol. 17, No. 15 (1989) pp. 5961-5971.
- Cotton G.H., (1989) "Detection of Single Base Changes in Nucleic Acids", *Biochemical Journal* 263:1-10.
- Su et al. J. Biol Chem (1988) 6829-6835.
- McKay, J. Molec Biol (1981) 145: 471-488.
- Hochuli et al. Bio/Technology (1988) 1321-1325.
- Saihi et al. Proc Natl Acad Sci. USA, (1989) 86: 6230-6234.

U.S. Patent

May 12, 1998

Sheet 1 of 10

5,750,335



U.S. Patent

May 12, 1998

Sheet 2 of 10

5,750,335

PATIENT NUCLEIC
ACID FRAGMENTS

NUCLEIC ACID FRAGMENTS FROM
NORMAL GENE(S)

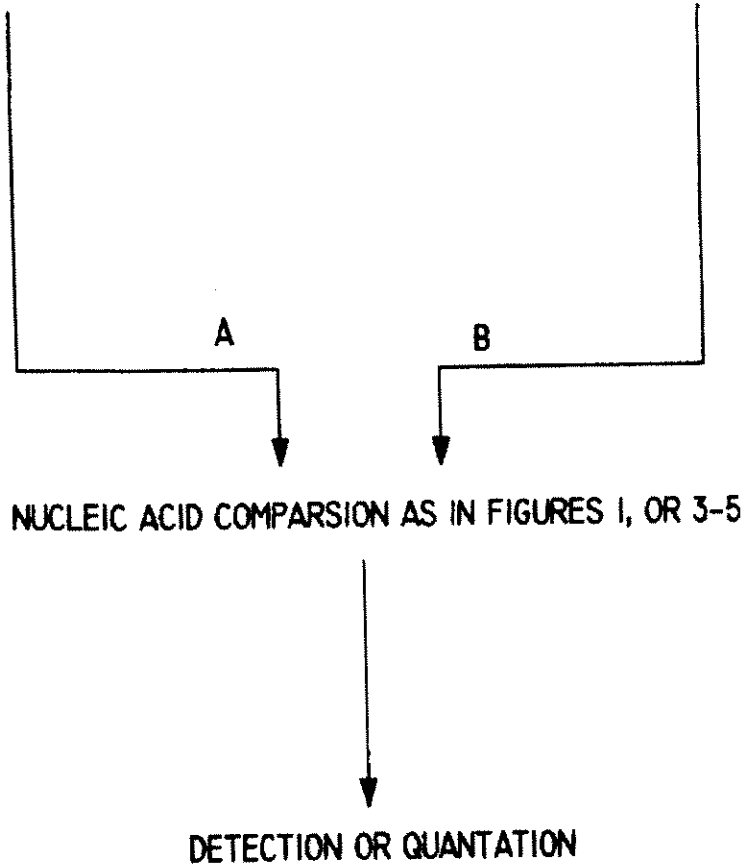


FIG. 2

U.S. Patent

May 12, 1998

Sheet 3 of 10

5,750,335

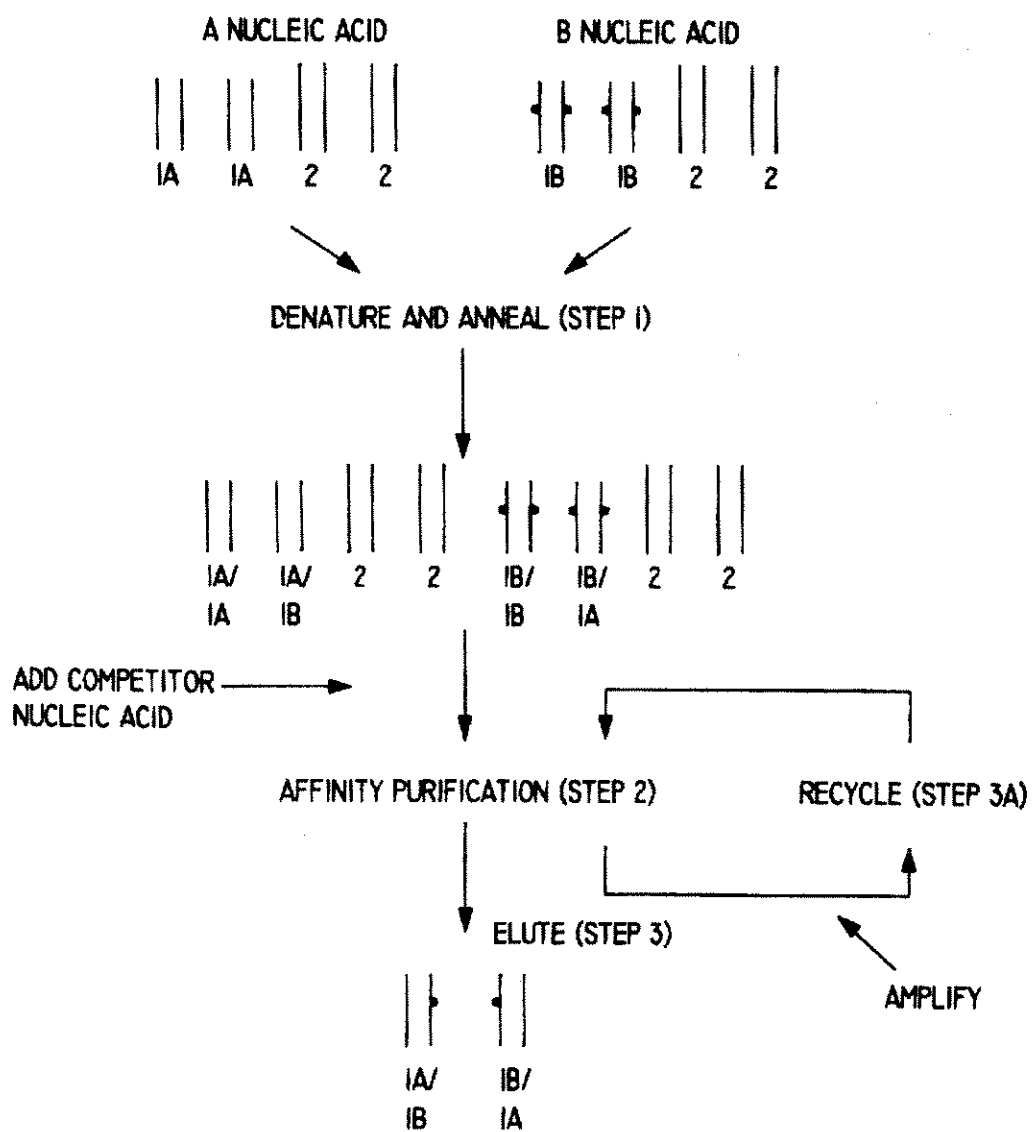


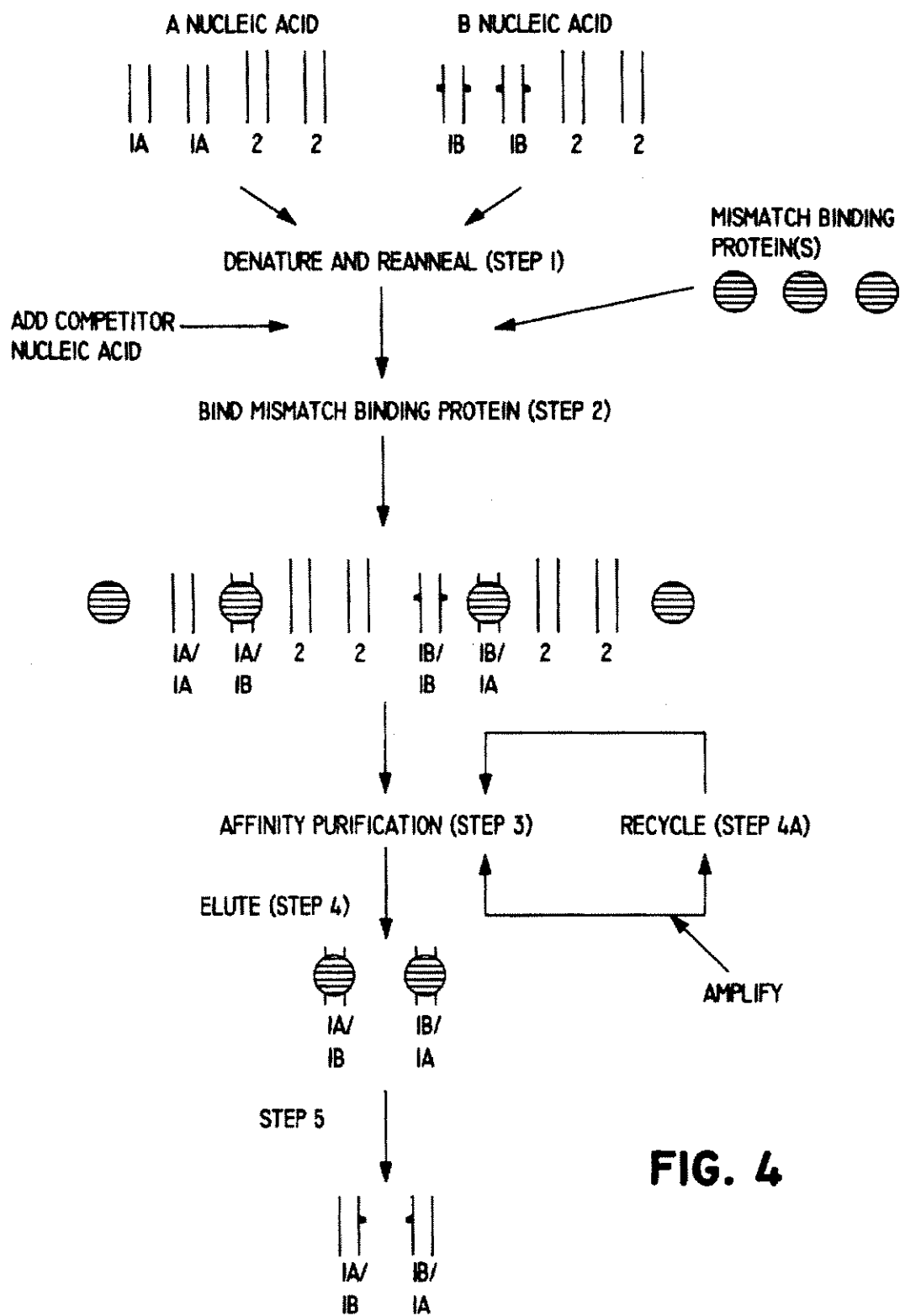
FIG. 3

U.S. Patent

May 12, 1998

Sheet 4 of 10

5,750,335

**FIG. 4**

U.S. Patent

May 12, 1998

Sheet 5 of 10

5,750,335

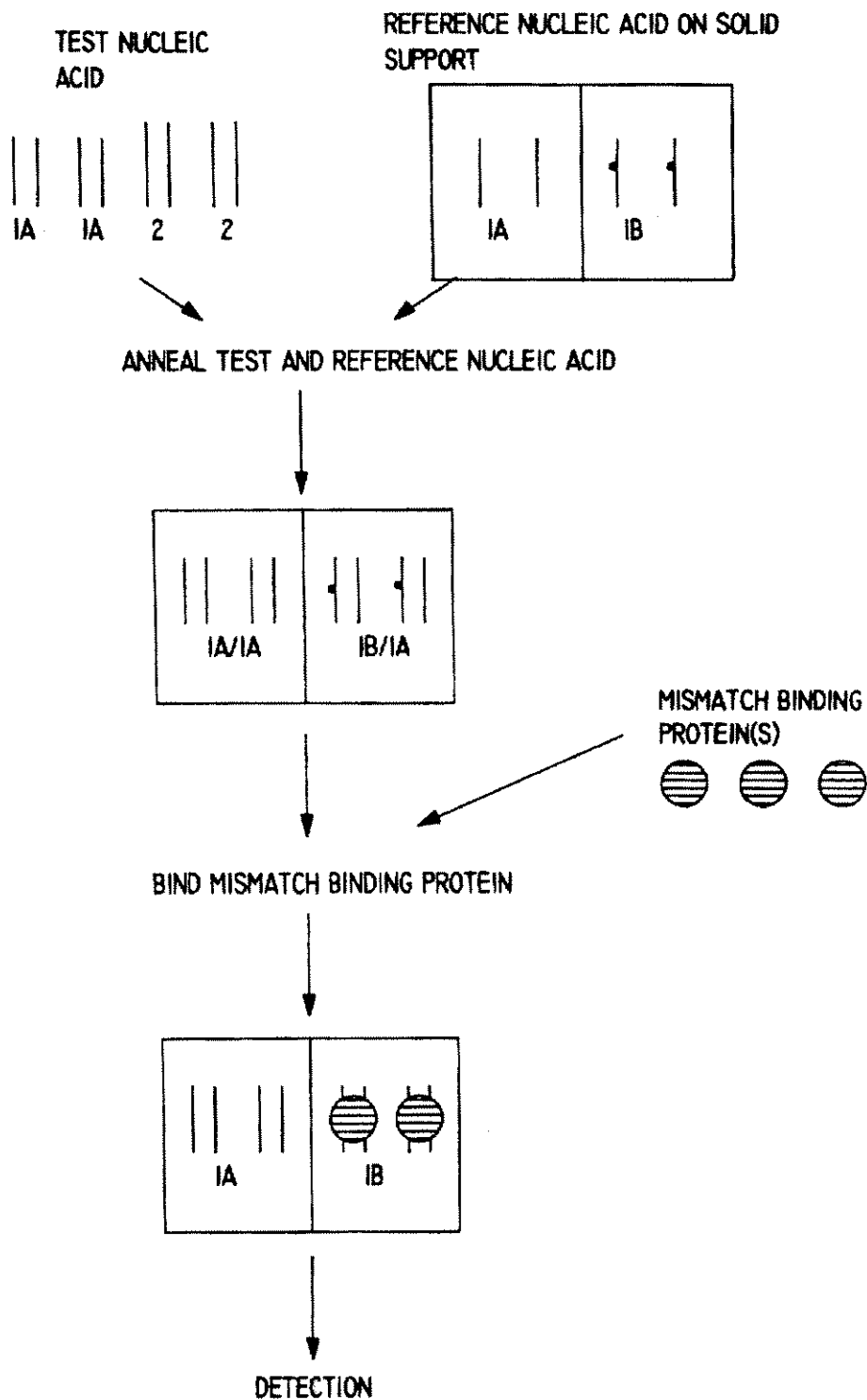


FIG. 5

U.S. Patent

May 12, 1998

Sheet 6 of 10

5,750,335

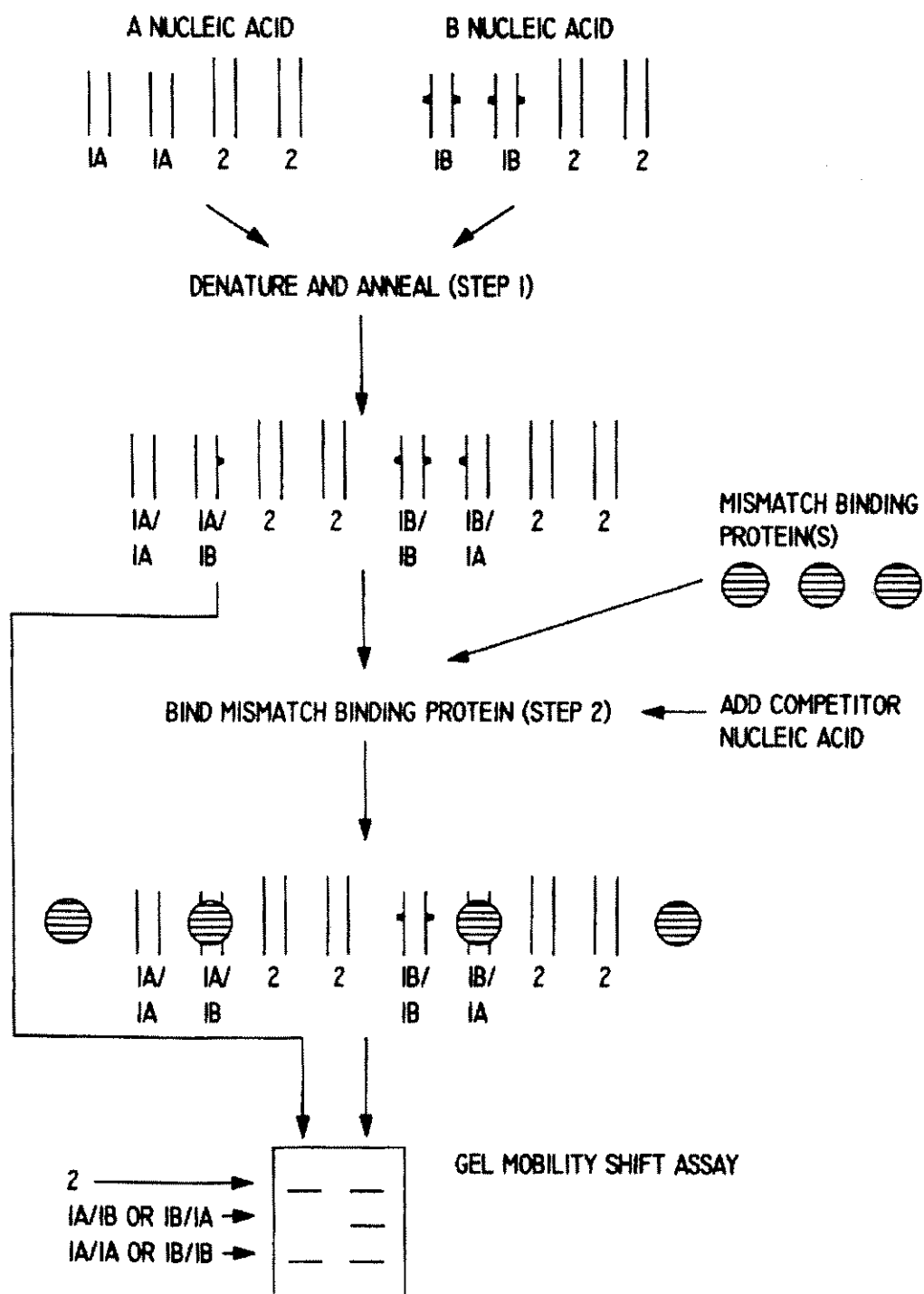


FIG. 6

U.S. Patent

May 12, 1998

Sheet 7 of 10

5,750,335



FIG. 7

U.S. Patent

May 12, 1998

Sheet 8 of 10

5,750,335

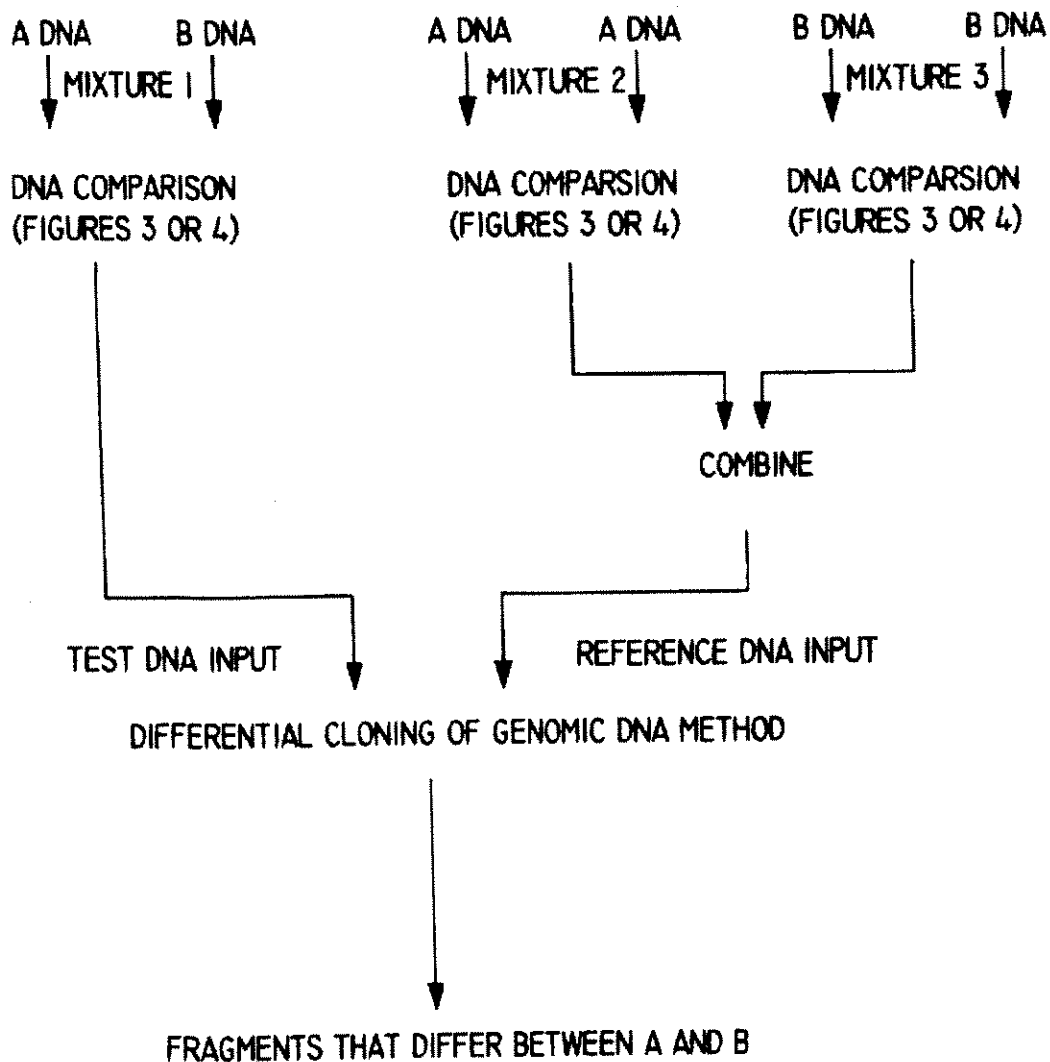


FIG. 8

U.S. Patent

May 12, 1998

Sheet 9 of 10

5,750,335

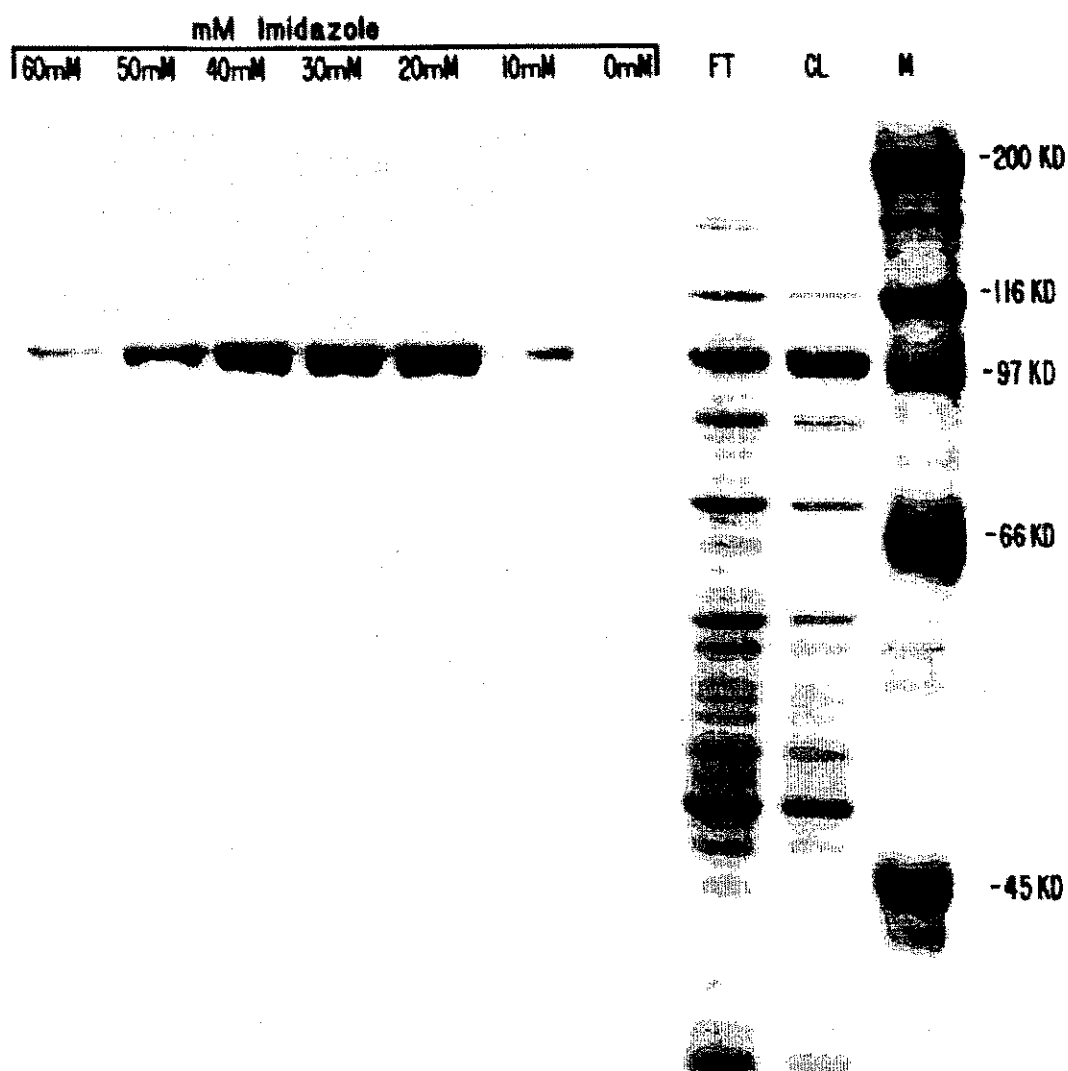


FIG. 9

U.S. Patent

May 12, 1998

Sheet 10 of 10

5,750,335

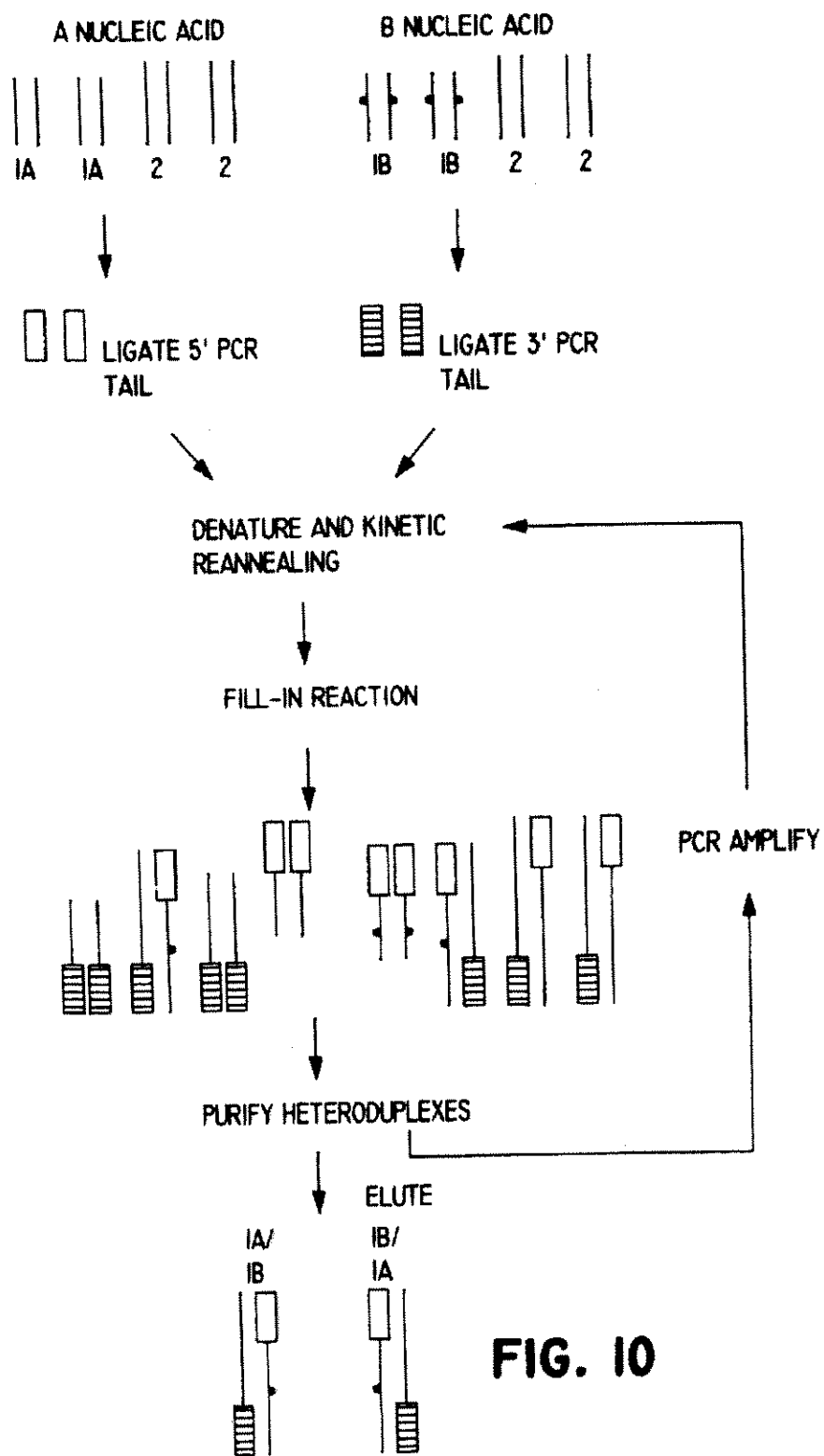


FIG. 10

5,750,335

1

SCREENING FOR GENETIC VARIATION

This application is a continuation-in-part of U.S. Ser. No. 07/874,192, filed Apr. 24, 1992, now abandoned.

The invention relates to the detection of sequence differences between test and reference nucleic acids; that is, to means and methods for the detection of the existence in a test polynucleotide of a genetic defect, or variation, from a reference, typically wild-type, polynucleotide. The invention is useful in clinical, forensic, and research contexts.

BACKGROUND OF THE INVENTION

Methods known in the art for comparing nucleotide sequence differences in DNA molecules are reviewed in Cotton, R., 1989, *Biochem. J.* 263:1, and include those aimed at detecting sequence differences when the sequence and location of a given region of DNA are known, discovering previously unknown mutations in a known region of DNA, and locating a previously unknown region containing a mutation.

Previous methods of detecting known sequence differences include: the failure of an oligonucleotide having a wild-type DNA sequence to hybridize under stringent conditions to sample DNA containing a mutation, the failure of PCR primers to hybridize under stringent conditions to sample DNA containing a mutation, and the consequent failure of sample DNA containing a mutation to become amplified using PCR; the failure of adjacent oligonucleotides to ligate due to a failure of one or both oligonucleotides to hybridize under stringent conditions to sample DNA containing a mutation; the use of primer extension analysis to detect incorporation of differentially labeled bases where the primer hybridizes to the sample DNA adjacent to the mutation; and the detection of changes in cleavability of a restriction enzyme site as an indicator of the presence of a mutation.

Previous methods of detecting a mutation of unknown identity within a known region of the genome include those in which a heteroduplex molecule is created from one strand of test DNA and one strand of reference DNA. Mismatches between the reference and test DNAs may be detected by carbodiimide modification of mismatched Thymidine (T) and Guanine (G) bases and detection of the resultant mobility shifts of modified versus control DNA; by ribonuclease cleavage of mismatched pyrimidine bases of RNA/DNA hybrids, and detection of points of cleavage in the molecule; by detection of differences in melting temperature between heteroduplex and homoduplex DNA, e.g., by denaturing gel electrophoresis; and chemical modification and cleavage of mismatched bases using hydroxylamine (to modify cytosine) or osmium tetroxide (to modify thymidine) modification and piperidine cleavage, and subsequent detection of cleaved DNA. Additional methods for detecting an unknown mutation within a region of DNA include: detecting differences in secondary structure by looking for differential mobility in gels of single stranded reference and test DNA; and by direct sequencing of both reference and test DNAs.

Several methods of locating mutations where both the identity and region of the mutation are described in the art. RFLP analysis, in which Restriction Fragment Length Polymorphisms are analyzed, identifies sequence differences which occur at restriction enzyme cleavage sites of test and reference DNAs, or by the insertion or deletion of a number of bases. RFMP analysis (Gray, 1992, *Amer. J. Hum. Genet.* 50:331) is a variation of RFLP analysis in which denaturing

2

gradient gel electrophoresis is used to identify sequence variations both at and between restriction enzyme cleavage sites.

The Southern Cross method, described in Potter and Dressler (1986, *Gene* 48:229), also depends upon sequence differences between test and reference DNAs that occur at sites of restriction enzyme cleavage. In this method, a reference DNA is digested with one or more restriction enzymes and analyzed by a modified Southern procedure. According to this modified Southern procedure, hybridization of two identical membranes, which are positioned at 90° angles with respect to each other, gives a signal that forms along a diagonal line of hybridization. In contrast, where test and reference membranes are hybridized at 90° angles, differences in restriction fragment patterns between the test and reference DNAs are indicated by off-diagonal signals.

Finally, the differential genomic DNA cloning method depends upon the inability of dephosphorylated reference DNA in a reference/test DNA hybrid to ligate to dephosphorylated vector DNA. In this method, described in Yokata and Oishi (1990, *Proc. Nat. Aca. Sci.* 87:6398), test and reference DNAs are digested separately with restriction enzymes, reference DNA is then dephosphorylated, and the two DNAs are combined at a ratio of 100/1 of reference to test DNA. The mixture is subjected to agarose gel electrophoresis, and the DNA is denatured and renatured in the gel, such that unique restriction fragments will likely self-anneal and non-unique fragments will likely reanneal with reference strands. Subsequent cloning of the reannealed fragments will favor reannealed test DNA clones, since the dephosphorylated reference DNA or reference/test hybrids will not be ligated to a dephosphorylated vector.

DNA mispairing can occur in vivo and is recognized and corrected by repair proteins. Mismatch repair has been studied most intensively in *E. coli*, *Salmonella typhimurium*, and *S. pneumoniae*. The MutS, MutH and MutL proteins of *E. coli* are involved in the repair of DNA mismatches, as is the product of the *uvrD* gene in *E. coli*, helicase II. MutS appears to play a central role in mismatch correction. Besides the repair system directed by Dam-mediated methylation of d(GATC) sites, MutS is also active in two other less efficient mismatch repair processes. One of these processes acts on symmetrically methylated DNA and may serve to repair mismatches produced during recombination. The other corrects cytosine (C) to Thymidine (T) transitions at the internal C of the Dcm methylase sequence d(CCA/TGG) or subsets thereof and also requires *mutL*⁺ and *dcm*⁺.

Mismatched base pairs can arise in vivo during homologous recombination of allelic genes, by chemical modification of DNA, or from errors made by DNA polymerase. Repair of mismatched DNA base pairs has been invoked to explain a variety of genetic phenomena, including gene conversion in *Neurospora* spp. and other fungi (Mitchell, 1955, *Proc. Nat. Aca. Sci.* 41:215; Rossignol, 1969, *Genetics* 63:795), postmeiotic segregation in *Saccharomyces cerevisiae* (Williamson et al., 1985, *Genetics* 110:609), high negative interference and gene conversion in lambda phage crosses (Nevers et al., 1975, *Mol. Gen. Genet.* 139:233; White et al., 1974, *Proc. Nat. Aca. Sci.* 71:1544; Wildenberg et al., 1975, *Proc. Nat. Aca. Sci.* 72:2202), and the existence of high and low efficiency transforming markers in *Streptococcus pneumoniae* (Ephrussi et al., 1966, *J. Gen. Physiol.* 49:211; Lacks, 1966, *Genetics* 53:207).

Jiricny et al. (1988, *Nucl. Ac. Res.* 16:7843) performed in vitro binding experiments using MutS and a series of synthetic DNA duplexes containing known mismatches or

5,750,335

3

mismatch analogues of the purine/pyrimidine type in order to demonstrate that MutS binds in vitro to double-stranded DNA containing a mismatched nucleotide pair. Su et al. (1986, *Proc. Nat. Aca. Sci.* 83:5057) have shown that highly purified MutS binds to a purified 120 base pair restriction fragment containing a single mismatch in vitro and protects approximately 22 nucleotides surrounding the mismatch against DNase attack. Su et al. (1988, *J. Biol. Chem.* 263:6829) demonstrates that MutS recognizes all eight possible DNA base mismatches.

McKay (1981, *J. Mol. Biol.* 145:471, hereby incorporated by reference), describes a method of purifying certain SV40 DNA restriction fragments using an immunoprecipitation procedure in which the SV40 T antigen-related protein binds to these DNA fragments. Blackwell and Weintraub (1990, *Science* 250:1104), hereby incorporated by reference, describes a method of purifying DNA sequences that bind to a protein of interest based on amplification of a binding site. The protein of interest is bound to DNA fragments and the bound fragment(s) is isolated using an electrophoretic mobility shift assay.

Objects of the invention include methods for rapid and accurate genetic screening and diagnosis by comparing two nucleic acids for differences in their nucleotide sequences. Another object is to diagnose genetic diseases in mammals, especially humans, by rapid screening for a previously observed mutation(s) known to cause a genetic disease. Another object is to rapidly screen the genome of an individual for genetic variation of a specific region of DNA, where the nature and position of the variation is unknown, by comparing a nucleic acid sequence known to reflect normal gene function with a nucleic acid sample suspected to contain a genetic defect. Yet another object is to locate previously unknown mutations of a nucleotide sequence and to identify the sequence itself, where the nature and position of the mutation within a region of the genome is unknown, and where the location of the region itself is unknown.

SUMMARY OF THE INVENTION

The invention provides methods of detecting and/or identifying polynucleotide sequence differences which may be the basis for genetic disease. The method involves hybridizing a "test", i.e., a potential variant, nucleic acid, e.g., from a patient, with a nucleic acid standard. If the test and standard (reference) nucleic acids contain one or more nucleotide sequence differences, then the double stranded nucleic acid formed from hybridization of the sequences will contain one or more nucleotide pair mismatches, i.e., will comprise a heteroduplex. In accordance with the invention, protocols are provided which permit detection of the presence of the heteroduplex, and/or segregation of a fraction rich in heteroduplex. The detection and fractionation methods involve exploitation of the selective binding properties of mismatch binding proteins.

The invention encompasses methods which allow for detection of differences between nucleotide sequences with greatly increased sensitivity. The methods of the invention allow one to detect single or multiple nucleotide differences between a nucleic acid standard and a sample nucleic acid without relying on restriction fragment length differences. The invention also provides for enrichment of heteroduplex fragments containing mismatches, even in a sample containing excess homoduplex, thereby achieving more sensitive detection of the mismatch. The methods also may be used quantitatively to determine the fraction of heteroduplex fragments in a mixture, and the proportion of mismatch

4

binding protein bound to heteroduplex, and thus also may be used to determine the number of mismatches within a test sample. The methods also allow for recovery of nucleic acid fragments containing sequence mismatches from a mixture containing excess fully complementary fragments. Recovered fragments may be analyzed further, for example, to determine the identity and position of the mismatch by determining the nucleotide sequence of the mismatch region.

In a first aspect, the invention features methods of genetic screening for a nucleotide variation which generally include the following steps. A mixture of nucleic acids which includes heteroduplex nucleic acids, i.e., heteroduplex including a test nucleic acid strand hybridized with a reference nucleic acid strand generated by annealing test and reference nucleic acid, and which includes a mismatched nucleotide pair, is subjected to a mismatch binding protein under conditions which promote binding of the protein to heteroduplex in the mixture to form a heteroduplex/protein complex. The presence of the mismatched nucleotide pair then is detected, using the methods disclosed below, as an indication of the presence of genetic variation between the test and reference nucleic acids.

In preferred embodiments of this aspect of the invention, the mixture provided may be a complex mixture of different nucleic acid fragments, some of which are heteroduplex fragments, but many or a majority of which are homoduplex nucleic acids. The test nucleic acid may be isolated from a collection of organisms and may include nucleic acid from any tissue or cell of several members of a species. Alternatively, the test nucleic acid may be sampled from an individual and thus may comprise nucleic acid from one unique representative of a species. In addition, the test nucleic acid may be suspected, but not known, to contain a nucleotide variation from a wild-type sequence which encodes a normal, functional protein or regulatory element. A nucleotide variation in the test nucleic acid comprises one half of a mismatched nucleotide pair when the test nucleic acid is hybridized to the reference nucleic acid.

The mixture of nucleic acids provided in the method typically are generated by annealing the test and reference nucleic acids. The test nucleic acid may be produced by cleaving double stranded test nucleic acid into a fragment which spans the same nucleotide region(s) as the reference nucleic acid(s). Both the test and reference nucleic acids may be either single or double stranded. If either is double stranded, the test mixture must be "melted", i.e., denatured to produce single stranded polynucleotide, before annealing. Generally, the test and the reference nucleic acids may be genomic DNA, cDNA, mRNA, synthetic polynucleotide, mitochondrial DNA, amplified or circular DNA, or other single or double stranded polynucleotide, from whatever source. While it is preferable that the reference nucleic acid be single stranded, it also may be double stranded.

The annealed mixture of test and reference nucleic acids will include a concentration of heteroduplexes if this test nucleic acid embodies at least one base difference from the reference. The heteroduplexes present in this mixture may be fractionated from the mixture by affinity purification in which a mismatch binding protein binds to the heteroduplexes preferentially to the homoduplexes in the mixture. The bound heteroduplexes may then be recovered from the affinity purification, e.g., released, to produce a fraction which contains a higher concentration of heteroduplex.

The methods of genetic screening also may include the immobilization of reference nucleic acid to a solid support.

5,750,335

5

For example, reference nucleic acids may be immobilized to a solid surface in an array of plural, spaced-apart spots. The spots of reference nucleic acid are then exposed separately under hybridizing conditions to a test nucleic acid such that the test and immobilized reference nucleic acids are able to form a hybrid. The hybrids then are contacted with the mismatch binding protein under conditions sufficient to allow the binding protein to bind to a heteroduplex containing a mismatched nucleotide pair. Finally, the bound mismatch binding protein, or the heteroduplex/protein complex, is detected as an indication of genetic variation between the test sample and the reference nucleic acid at that spot.

Detection of the heteroduplex may be conducted by detecting the mismatch binding protein that is bound to the heteroduplex, e.g., using a labeled form of the mismatch binding protein or a separate binding protein such as an antibody specific for the mismatch binding protein. Alternatively, the heteroduplex may be detected by detecting the complex, e.g., with an antibody specific for an epitope on the heteroduplex/mismatch binding protein complex. Alternatively, the bound mismatch binding protein or bound heteroduplex may be released from the complex before detection of the released component. Alternatively, the mismatch binding protein may modify the heteroduplex before it releases, and the modification may be subsequently detected. The heteroduplex itself can include a detectable moiety, e.g., a radioactive or other label bound to the reference nucleic acid, and the detecting step can include detecting the detectable moiety after fractionation of the heteroduplex. The methods may also include, in addition to detecting the presence of a mismatched nucleotide pair, determining the identity or location of the nucleotide variation in the test strand. The identity or location of the nucleotide variation may be determined by analyzing the nucleotide sequence of the test nucleic acid strand and comparing it to the sequence of the reference strand.

In a second aspect, the invention features methods of selectively enriching a nucleic acid preparation in fragments containing a nucleotide variation, by enriching for heteroduplex nucleic acids in a mixture. Selective heteroduplex enrichment of a mixture which includes a first concentration of heteroduplex nucleic acids may be performed by separating the heteroduplex nucleic acids by affinity purification in which the mismatch binding protein binds to heteroduplex, and recovering heteroduplex to produce a mixture that contains a second, higher concentration of heteroduplex. As a variation on this method, the mixture first is reacted with a mismatch binding protein such that the heteroduplex binds to the protein to form a heteroduplex protein complex, and then the complex is separated from the mixture by affinity purification to produce a mixture having a higher concentration of heteroduplex. In both variations of this aspect of the invention, the affinity purification step involves a binding reaction in which the heteroduplex is selectively bound by a mismatch binding protein which preferably is coupled to a solid support, followed by elution. The binding and elution steps may be repeated interactively until a desired degree of purification of heteroduplexes is achieved. Numerous modifications of this general procedure are encompassed by the invention. For example, the mismatch binding protein/heteroduplex complex may be bound by 1) a protein specific for one or both components of the complex, e.g., an antibody, 2) a metal column capable of binding to a histidine tail engineered onto the mismatch binding protein, or 3) a protein capable of binding to a flag sequence on the mismatch binding protein. A solid support may not be preferable; e.g., an antibody may be used to

6

immunoprecipitate the mismatch binding protein/heteroduplex complex.

In both aspects of the invention, the test nucleic acids may be prepared by, for example, performing a polymerase chain reaction on a region of interest in test nucleic acid sample. In addition, an amplification step, e.g., by polymerase chain reaction, may be useful at other points of the methods, e.g., after affinity purification of heteroduplex nucleic acids to produce an amplified heteroduplex sample. Where a PCR step is performed, it may be necessary to ligate PCR tails to the test, reference, or heteroduplex nucleic acids prior to the mismatch binding protein binding reaction.

In both aspects of the invention, when the reference nucleic acid is labeled, the methods may include the additional step of adding excess unlabeled nucleic acid to the mixture of test and reference nucleic acids to serve as a competitor to mismatch binding protein binding, thereby to reduce background. Background may be caused by the nonspecific binding of mismatch binding protein to homoduplex nucleic acid. In this case, detection of labeled reference nucleic acid does not correlate directly with the amount of heteroduplex present, even though purification was conducted with mismatch binding protein because of nonspecific interactions between the mismatch binding protein and homoduplex nucleic acid. However, the presence of unlabeled competitor creates a dilution effect on labeled homoduplex nucleic acid, formed by annealing of reference/reference strands or test/test strands, which otherwise would be mistaken for heteroduplex. Alternatively, background may be reduced using an amplification step. PCR tails are ligated to the test and reference nucleic acids but not to the competitor nucleic acid. Excess competitor is added to the mixture prior to binding of mismatch binding protein. The subsequent amplification of presumed heteroduplex nucleic acid purified from the complex also will result in amplification of nonspecifically bound homoduplex nucleic acid. However, the presence of excess competitor nucleic acid lacking PCR tails will dilute out the effect of nonspecific binding because nonspecifically bound competitor nucleic acid will not be amplifiable.

In another aspect, the invention features apparatus for conducting comparisons of the sequence of test and reference nucleic acid, and for determining the existence or nature of a difference between two or more nucleic acid sequences. Broadly, these apparatus include, as essential elements, a mismatch binding protein, and either or both means for detecting the presence of the protein or a protein/heteroduplex complex, and/or means for separating heteroduplex from homoduplex in a mixture.

A kit for detecting a heteroduplex nucleic acid as an indication of genetic variation may include an array of separately spaced reference nucleic acids coupled to a support, and a mismatch binding protein. Preferably, the mismatch binding protein is labeled, but alternatively, the kit may include a protein that binds the mismatch binding protein, e.g., a labeled protein such as an antibody or an unlabeled antibody that is bound by a labeled antibody. The protein capable of binding the mismatch binding protein may be immobilized on a solid support.

A detection kit may also include a mismatch binding protein immobilized on a solid support, and means for detecting a heteroduplex bound to the support through the protein, or eluted from the support.

The invention also features a kit for separating a heteroduplex nucleic acid from a mixture of heteroduplex and homoduplex nucleic acids, which includes a mismatch bind-

5,750,335

7

ing protein, a moiety capable of binding a mismatch binding protein, or a moiety capable of binding a complex comprising a mismatch binding protein and a heteroduplex, all coupled to a solid support, and means for separating the heteroduplex from homoduplex. Any of the kits may include a reference nucleic acid.

In still another aspect, the invention features a solid support, e.g., an affinity matrix for binding heteroduplex nucleic acids. The support comprises a mismatch binding protein coupled to a high surface area matrix. Alternatively, the support may comprise immobilized moieties which bind a mismatch binding protein, or bind a heteroduplex/mismatch binding protein complex.

As used herein, a "mismatch binding protein" refers to any organic moiety, e.g., a protein, polypeptide, organic analog thereon, or other moiety or mixture of moieties, which bind preferentially to regions of double-stranded nucleic acids containing a mismatch. The mismatched regions may be as little as one nucleotide pair and may be as large as 5-10 nucleotide pairs, e.g., a small loop region. Such binding proteins include but are not limited to naturally occurring proteins, such as MutS, MutL, MutH, and MutU (helicase II) from *E. coli* and *Salmonella typhimurium*, HexA and HexB from *S. pneumoniae*, and mismatch binding proteins found in higher organisms, including humans (Jiricny et al., 1988 Proc. Nat. Aca. Sci. U.S.A. 85:8860; Stephenson et al., 1989, J. Biol. Chem. 264:21177), and analogs thereof which contain amino acid differences that do not destroy binding of the protein to the mismatched nucleotides, but may have properties not present in conventional mismatch binding protein, e.g., thermostability. As used herein, "mismatch binding protein" also includes proteins which do not naturally bind a nucleotide mismatch, but which has been altered or engineered to bind a nucleic acid fragment containing mismatched nucleotides, and muteins; derivatives, truncated analogs, or species variants of naturally occurring mismatch binding proteins. The definition also includes an antibody or a mixture of antibodies that recognizes and binds heteroduplex nucleic acids. Also included in the invention are mismatch binding proteins that modify nucleic acids containing mismatches, thus allowing the nucleic acid to be subsequently recognized by other proteins or means.

As used herein, "homoduplex" refers to double stranded nucleic acid containing first and second strands which are fully complementary. "Heteroduplex" refers to double stranded nucleic acid containing first and second strands which are substantially complementary, but which contains regions of noncomplementarity, i.e., one or more mismatched nucleotide pairs. Regions of noncomplementarity may cause small loops to form within one strand of the heteroduplex. There may be as few as one region of noncomplementarity per heteroduplex, or many regions, so long as the heteroduplex can form a stable hybrid under conditions selected to-form the hybrid. A non-complementary region may include insertions or deletions of one or more bases of one strand relative to the other strand. "Competitor" nucleic acid refers to homoduplex nucleic acid that is either unlabeled or does not contain PCR tails, or that is distinguishable from heteroduplex nucleic acid. "Excess homoduplex" nucleic acid refers to a mixture containing at least two-fold, preferably at least five- or ten-fold, and most preferably at least 100-fold more homoduplex nucleic acid than heteroduplex nucleic acid, where the excess homoduplex nucleic acid is a natural by-product of the process that created the heteroduplex nucleic acid. "Excess competitor" nucleic acid refers to a mixture containing at least two-fold, preferably at least five-

8

or ten-fold, most preferably at least 100-fold more competitor homoduplex-nucleic acid than heteroduplex nucleic acid. "Nucleic acid" refers to DNA or RNA containing naturally occurring nucleotides or synthetic substitutions thereof. "Test" nucleic acid refers to single- or double-stranded DNA or RNA to be compared to the nucleic acid standard, e.g., DNA from a patient suspect of having a genetic disease. "Reference nucleic acid" refers to a single or double-stranded nucleic acid standard, e.g., a nucleic acid encoding a normal protein or regulatory function. "Mismatched nucleotide pair" refers to a nucleotide pair which does not match according to Watson/Crick base pairing, i.e., is not G:C, A:T, or A:U. A "nucleotide variation" is a nucleotide sequence difference between a test nucleic acid and a reference nucleic acid, and constitutes as little as one base pair of a mismatched nucleotide pair. "Amplify" means to make multiple copies of a nucleic acid fragment or a mixture of nucleic acids. "PCR" means polymerase chain reaction, and "PCR tail" refers to oligonucleotide duplexes which are ligated to the ends of nucleic acids and which, upon denaturation, may hybridize to complementary primers used to prime the synthesis of DNA. "Labeled" means containing a detectable moiety or a moiety which participates in a reactions resulting in detection, e.g., a chromogenic reaction. A detectable moiety may, include but is not limited to a radioactive marker, e.g., ³²P, and non-radioactive markers, e.g., biotin. "Affinity purification" or "affinity fractionation" means to separate heteroduplex or heteroduplex/binding protein complex from other components based on the affinity of the heteroduplex or complex. An "affinity matrix" is a solid support which is used to affinity purify heteroduplex or heteroduplex/binding protein complex.

As used herein, a nucleic acid "isolated from an organism" refers to DNA or RNA that has been extracted directly from cells or tissue of one or more members of a species, e.g., prokaryotic, eukaryotic, or mammalian, especially human DNA or RNA from human cells or tissue; or to DNA that has been cloned from genomic DNA or from RNA sequences; or to DNA that has been amplified from an organism's DNA using the technique of polymerase chain reaction. Nucleic acid "native, to an individual" refers to DNA or RNA that has been extracted from, cloned from, or amplified from cells or tissue of a member of a species. Where a nucleic acid is "suspected to contain" a nucleotide variation, it is not known whether the nucleic acid contains the variation prior to performing the method of the invention.

Other features and advantages of the invention will be apparent from the following description of the preferred embodiments, from the drawing, and from the claims.

DETAILED DESCRIPTION OF THE INVENTION

We first briefly describe the drawings.

Drawings

FIG. 1 schematically illustrates a method of detecting nucleic acid sequence mismatches;

FIG. 2 schematically illustrates a method for performing genetic disease diagnosis using a method of the invention in which the reference nucleic acid is labeled or detected using other means;

FIG. 3 schematically illustrates a method of affinity purifying heteroduplex nucleic acid molecules using a mismatch binding protein;

FIG. 4 schematically illustrates heteroduplex affinity purification in which heteroduplex mismatch binding protein complexes are fractionated;

5,750,335

9

FIG. 5 schematically illustrates a method of detecting nucleic acid sequence mismatches using an array of plural, separate reference nucleic acids arranged on a solid support;

FIG. 6 schematically illustrates a method of detecting nucleic acid sequence mismatches using a band shift assay;

FIG. 7 illustrates the results of a band shift assay; and
FIG. 8 schematically illustrates a method of differentially cloning nucleic acids sequences containing sequence variations.

FIG. 9 is a polyacrylamide gel showing the results of purification of histidine-tagged MutS.

FIG. 10 schematically illustrates a method of differentially analysing test/reference nucleic acid hybrids containing a mismatch.

We next describe preferred embodiments of the invention.

I. Preparation of Nucleic Acids

Test or reference nucleic acids can be prepared using a variety of techniques. For example, nucleic acid can be extracted from cells and used directly, or a specific region of extracted nucleic acid may be amplified; alternatively, nucleic acid may be synthesized.

Cultured cells, tissue or blood samples may be used as a source or as the source of a nucleic acid sequence. Cultured monoclonal cell lines will give a single type of test nucleic acid, and cultured polyclonal cell lines can be used to check for differences between one standard nucleic acid and a library of nucleic acids containing many different test DNAs. Either chromosomal and/or extra-chromosomal DNA, such as plasmid DNA, can be isolated for use as test or reference nucleic acid.

Nucleic acid can be extracted from cells, purified, and digested with restriction enzyme(s) to create nucleic acid fragments, and also may be subsequently amplified. The polymerase chain reaction (PCR) can be used to amplify a given region of nucleic acid in order to limit the scope of inquiry to this region, by choosing appropriate primers that flank the region of interest. In addition, multiple primers can be used at once to amplify a set of regions of interest for simultaneous comparison.

Test or reference nucleic acid may also be prepared from synthetic DNA. DNA can be synthesized, and one or more oligonucleotides may be used as a test or reference nucleic acid. Oligonucleotides are particularly useful as reference nucleic acid for moderate size regions.

A test or reference nucleic acid may also include a mixture of two or more of cellular DNA, amplified DNA, and/or synthetic DNA, for simultaneous comparison of different nucleic acid loci.

1. Representational Difference Analysis.

If desired, a nucleic acid sample may be treated so as to reduce the complexity of the sample by removing irrelevant or unnecessary nucleic acid sequences, e.g., using representational difference analysis, subtractive hybridization or kinetic enrichment (Kinzler et al., *Nucleic Acid Research* 17, 10:3645 1989); Lisitsyn et al., *Science* 259:956 (1993), both references of which are hereby incorporated by reference). The complexity of a nucleic acid sample may be decreased significantly by preparing a representative portion of each of the test and reference nucleic acid samples, or of the denatured and reannealed test/reference sample, as described by Lisitsyn et al., *supra*. Nucleic acid populations of reduced complexity, i.e., "representations", allow for detection of nucleotide sequence differences between two complex genomes. One method of creating a representative portion of a nucleic acid sample is to selectively amplify certain fragments relative to others. For example, test or reference nucleic acid is first cleaved into restriction

10

fragments, and then PCR tails are ligated onto the ends of the fragments. If the restriction sites chosen for cleavage occur infrequently, then the average restriction fragment size will be large. Upon amplification of the tailed fragments using PCR primers that are complementary to the tail sequences, the smaller fragments of the mixture will be selectively amplified. Thus, a representative nucleic acid sample is created which contains the relevant sequences but is significantly less complex than the original nucleic acid sample. Subsequent reiterations of the method will further enrich the sample for relevant sequences.

Test or reference nucleic acids also may have identical primer sequences incorporated at their ends to permit the later amplification of the heteroduplex nucleic acid; for example, PCR tails may be added onto the ends of, e.g., the "A" and "B" samples in FIG. 1, prior to step 1, and PCR amplification may be performed at a later step in the procedure.

2. Differential PCR Tailing.

PCR also can be used so as to allow subsequent amplification of only test-reference hybrids, and thus reduce the frequency of test-test and/or reference-reference hybrids in the sample. FIG. 10 schematically illustrates this method. It will be appreciated that complete or partial digestion by multiple restriction enzymes yields non-symmetric 5' and 3' ends suitable for differential PCR tail ligation. Of course, the first PCR tail may be ligated onto reference nucleic acid and the second PCR tail may be ligated onto test nucleic acid. According to this method of the invention, only test-reference hybrids will undergo exponential amplification. This method is described in detail below.

3. Differential Strand Labeling.

Test and reference nucleic acids may also be differentially labeled to allow their progress to be traced through the comparison process. For example, a test nucleic acid can be left unlabeled and the reference nucleic acid (or another test nucleic acid) can be, for example, end-labeled with ³²P by a kinasing reaction. Any appropriate labeling method may be used; e.g., to permit detection of radioactively-labeled nucleic acid or chromogenic or chemiluminescent detection of, for example, a biotin labeled nucleic acid. In addition, determining the presence or absence of specific nucleic acid sequences may be achieved by differential detection, e.g., using different PCR primer sequences which are sequence specific for the fragments of interest. The subsequent selection of corresponding primer oligonucleotides for use in the PCR amplification reaction, followed by analysis of the amplified nucleic acid, will give amplification of the selected nucleic acid.

II. Preparation of Heteroduplexes and Homoduplexes

Heteroduplex nucleic acid includes double stranded nucleic acids in which the molecules contain one strand each from the test and reference nucleic acids. If the test and reference nucleic acids contain differences, annealing of test and reference strands will create heteroduplex molecules. Where the test and reference nucleic acids are completely homologous or the test and reference strands anneal as test/test or reference/reference hybrids, a homoduplex will be created. The heteroduplex molecule forms despite the mismatch because the remainder of the matched base pairs stabilizes the heteroduplex molecule. Thus, heteroduplex molecules are formed by fragments that are similar enough to anneal but that contain mismatches.

The degree of similarity necessary for a heteroduplex to be formed can be controlled by the stringency of the annealing conditions. For example, if the annealing reaction is run at an elevated temperature, single stranded molecules

5,750,335

11

will need to have increased sequence similarities before they can form heteroduplexes. Conditions for annealing of nucleic acids to form hybrids are well-known in the art or, if unknown, can be determined by routine experimentation. See, for example, Alt et al. (1978, *J. Biol. Chem.* 253:1357, hereby incorporated by reference).

A standard method of denaturing and reannealing nucleic acids which may be used to prepare heteroduplexes according to the invention is the following. The test nucleic acid is suspended in 100 μ l of 1 \times SSC buffer (0.15M NaCl, 0.015M Nacitrate) in an eppendorf tube. The tube is placed in a beaker of water, and the beaker of water is placed in a boiling water bath until the water in the beaker boils. After ten minutes of boiling, the beaker is removed from the water bath, and allowed to cool to 65° C., and placed in a 65° C. water bath. The 65° C. water bath is switched off. The nucleic acid is allowed to anneal during cooling of the 65° C. water bath to room temperature. The nucleic acid can then be ethanol precipitated and resuspended in TE buffer.

III. Identification of Heteroduplex Fragments

FIGS. 1-6 and 8 schematically illustrate methods for the detection and/or analysis of genetic differences according to the invention. FIG. 7 shows the results of one such identification.

In FIG. 1, a method of detecting a nucleotide pair mismatch is shown schematically. In step 1, test and reference nucleic acids (samples A and B, respectively, each sample containing two different nucleic acid fragments, 1 and 2, respectively), are denatured and reannealed such that single stranded molecules from sample A nucleic acid and sample B nucleic acid reanneal to form duplexes. Fragment 2 in each of the test and reference samples is identical (i.e., contains no mismatches), and forms a homoduplex after the reannealing process. In contrast, fragment 1A differs from fragment 1B by only a single base pair mismatch. When a single strand of fragment 1A reanneals with a single strand of fragment 1B, a heteroduplex nucleic acid molecule forms ("1A/1B" in the figure) containing a mismatched base pair. This is shown schematically in FIG. 1 as the mixture of denatured and reannealed fragments between steps 1 and 2. Fragments 1A/1B and 1B/1A each contain a nucleotide pair mismatch, whereas fragments labeled "1A/1A", "1B/1B", and "2" are fully complementary. The mixture of fragments is then subjected to a binding reaction in which the mismatch binding protein is allowed to bind to fragments containing mismatches. The results of the binding reaction are shown schematically in step 2 of FIG. 1, in which the protein is shown bound to each of fragments "1A/1B," and "1B/1A" containing mismatches. In step 3, the mismatches are detected and/or quantitated. Examples of detection and quantitation of nucleotide pair mismatches are disclosed herein. Optional steps in the method shown in FIG. 1 and in other figures include the addition of competitor nucleic acid prior to binding of the mismatch binding protein to reduce nonspecific binding to matched nucleic acid, and thus reduce background; and the amplification of a sample containing heteroduplex nucleic acid at some step prior to detection or quantitation. These optional steps are discussed more fully below.

In FIG. 2, a quantitative method of genetic disease diagnosis according to the invention is schematically shown. Patient nucleic acid is prepared according to conventional techniques, and cleaved into restriction fragments. The nucleic acid standard, to which the patient nucleic acid is to be compared, contains "normal" nucleic acid fragments, i.e., nucleic acid fragments having a sequence known to reflect the normal gene functions. In this example, either the

12

nucleic acid standard is labeled or the mismatch binding protein is labeled. The two nucleic acid samples are then subjected to any one of the methods of the invention, including those illustrated in the figures. This step is referred to as "Nucleic Acid Comparison" in FIG. 2. The results of the nucleic acid comparison, i.e., the detection or isolation of hybrid nucleic acid fragments of patient/standard nucleic acid containing one or more nucleotide pair mismatches, may be subjected to quantitative analysis by quantitating the data present in both input and output samples.

In FIG. 3, a method of selectively enriching for nucleic acid hybrids containing mismatches is shown. In this figure, the affinity purification step involves the selectively sequestering of heteroduplex nucleic acid using a mismatch binding protein. Step 1 of FIG. 3 is similar to step 1 of FIG. 1, and involves the denaturation and annealing of a test and a reference nucleic acid sample (A and B, respectively). The mixture of annealed nucleic acid is shown, as in FIG. 1. The annealed mixture is then subjected to an affinity purification reaction in which heteroduplex nucleic acid is bound by a mismatch binding protein under appropriate binding conditions, as described herein. The affinity purification reaction may be an immunoprecipitation reaction in which the mismatch binding protein is allowed to bind to the nucleic acid, followed by immunoprecipitation using an antibody, as described below. Alternatively, the affinity purification reaction may include subjecting the annealed mixture to mismatch binding protein coupled to beads, e.g., in a free slurry or poured into a column matrix. The bound heteroduplex nucleic acid will become sequestered with the beads and will thus be separable from the unbound nucleic acid. After separation, the bound nucleic acid is eluted or released (Step 3). The mismatch binding protein may be attached to any solid support that will permit the separation of free nucleic acid from nucleic acid bound by the mismatch binding protein.

Affinity purification of heteroduplex nucleic acid may involve any of a number of affinity purification techniques, and is not limited to that discussed above. For example, as shown in FIG. 4, the affinity step may involve selectively sequestering of the entire heteroduplex/mismatch binding protein complex, rather than just the heteroduplex nucleic acid itself. Steps 1 and 2 of FIG. 4 are similar to steps 1 and 2 of FIG. 1, in which the annealed mixture is formed and subjected to a binding reaction in which mismatch binding protein binds to heteroduplex nucleic acid in the mixture. In step 3, the heteroduplex/binding protein complexes are selectively retained, e.g., by a matrix to which an antibody specific for the binding protein is coupled. The complexes may then be eluted (step 4), followed by isolation of the heteroduplex nucleic acid (step 5), e.g., by phenol extraction of protein and ethanol precipitation of nucleic acid.

FIG. 5 shows an alternative method of genetic disease screening and diagnosis in which nucleotide pair mismatches are detected in a simple assay. This method is a specific embodiment of that shown in FIG. 1, and involves a solid support in which quantities of reference nucleic acid are spotted onto a membrane in an ordered pattern. The standard (reference) and the patient (test) nucleic acids are then denatured and annealed according to conventional techniques. After the hybrids are allowed to form, the membrane is subjected to a binding reaction in which mismatch binding protein is allowed to bind to any heteroduplexes which may have formed. After unbound mismatch binding protein is washed off the membrane, the presence of bound mismatch binding protein is detected using any appropriate detection technique disclosed herein or known in the art.

5,750,335

13

An alternative to fixing the reference nucleic acid on a solid support is to fix the test nucleic acid on a solid support. The technique outlined in FIG. 5 can be applied to this alternative method, with the modification that reference nucleic acid is annealed to the fixed test nucleic acid. Methods of fixing test nucleic acid to a solid support include crosslinking, alkaline transfer to a membrane, or other techniques, as described in Ausubel et al., eds., 1992, current protocols in Molecular Biology, John Wiley & Sons, New York, also herein incorporated by reference. Alternatively, in situ hybridization, also as described in Ausubel, can be used to directly anneal reference nucleic acid to test nucleic acid that is contained in sectioned cells. Annealing can be optionally performed in the presence of competitor nucleic acid.

Another alternative method of genetic disease screening or diagnosis involves the detection of nucleotide pair mismatches using a band shift assay. FIG. 6 illustrates this method. In steps 1 and 2, the patient (test) nucleic acid is denatured and annealed to reference nucleic acid and allowed to bind to mismatch binding protein, as described in FIG. 1. The bound nucleic acid is then electrophoresed on an agarose gel. This method takes advantage of the decreased mobility of bound heteroduplexes relative to unbound hybrids in agarose. As shown schematically in FIG. 6, the control lane (left), in which the annealed fragments were not subjected to mismatch binding protein, contains only homoduplex fragment 2 (top) and 1A/1A, 1B/1B, or unbound heteroduplex 1A/1B or 1B/1A (bottom), whereas the experimental lane (right) contains both homoduplex bands (top and bottom) and the middle heteroduplex band (1A/1B or 1B/1A). The results of such an assay are shown in FIG. 7. Mismatch binding protein was allowed to bind under binding conditions to a mixture of nucleic acid fragments, and then subjected to agarose gel electrophoresis. The mobility of the nucleic acid fragment in the mixture that contained a nucleotide pair mismatch is near the top of the gel (lane 2) and thus was selectively slowed relative to the faster running unbound nucleic acid fragments, which migrated to the bottom of the gel. The control lanes in FIG. 7 (lane 1 and 3) show that when no mismatch binding protein is added to the binding reaction, there is no binding to fragments and consequently no fragments migrating with the bound fragments in the gel.

A genetic disease may be not only detected, but also further analyzed to learn more about the genetic cause of the disease using the mismatch detection and isolation methods of the invention. Such analysis may include determining the nucleotide sequence of the strands of the isolated heteroduplex nucleic acid, or may involve the cloning of that portion of the patient's nucleic acid that contains the nucleotide sequence difference. FIG. 8 schematically illustrates a method differential cloning of heteroduplex strands. Test nucleic acid includes heteroduplex nucleic acid from samples A and B as shown in FIGS. 3 or 4. This nucleic acid was prepared by annealing a patient and a standard nucleic acid and purifying the heteroduplexes bound by the mismatch binding protein to produce mixture 1 in the figure. Reference nucleic acid in FIG. 8 is prepared from mixtures 1 and 2. Mixture 2 is prepared by denaturing and annealing sample A with itself and purifying heteroduplexes bound by mismatch binding protein. Similarly, mixture 3 is prepared by denaturing and annealing sample B with itself and purifying heteroduplexes bound by mismatch binding protein. Mixtures 2 and 3 are then pooled without denaturing and reannealing again to produce the reference nucleic acid. The test A/B and reference A/A and B/B nucleic acids are then subjected to the differential cloning method described

14

below. This method produces clones of A and B nucleic acids that were part of a A/B heteroduplex.

IV. MutS Binding Reaction

The mismatch binding protein MutS from *Salmonella typhimurium* selectively binds mismatches in heteroduplex molecules. MutS also binds mismatches that include deleted or added bases. Additional mismatch binding factors, such as MutL, can also be used in the binding reaction as an alternative to or in combination with MutS, to increase binding. MutS protein can be purified using the MutS overproducer plasmid pGW1825 (Haber et al., 1988, J. Bacteriol. 170:197) and the method of Su and Modrich (1986, Proc. Nat. Aca. Sci. 83:5057). MutL has been cloned into plasmid pGW1842 (Mankovich et al., 1989, J. Bacteriol. 171:5325), and can be purified using the method of Griley et al. (1989, J. Biol. Chem. 264:1000). Haber et al., 1988, Su et al. 1986, Griley et al., 1989, and Mankovich et al. 1989 are all hereby incorporated by reference.

The mismatch binding protein/heteroduplex binding reaction is typically performed as follows. The reaction is performed in assay buffer (20 mM Tris.HCl pH 7.6, 5 mM MgCl₂, 0.1 mM DTT, and 0.01 mM EDTA) for 30 minutes on ice. Typical binding reactions are 10 pmol total volume, with 0.2 pmol of duplex DNA and 40 pmol of mismatch binding protein, e.g., MutS. The addition of ATP to the binding reaction may increase the efficiency of binding of the protein or of cofactors such as MutL.

In addition to selectively binding heteroduplex nucleic acid, MutS nonspecifically binds to homoduplex nucleic acid to some degree. In order to reduce nonspecific binding, competitor (i.e., homoduplex) nucleic acid may be added to the heteroduplex mixture prior to the binding reaction or the affinity fractionation step, as shown in FIGS. 1, 3, and 4. Where the test or reference nucleic acid is labeled, as shown in FIG. 2, the use of excess unlabeled competitor DNA will cause most non-specific binding to occur on unlabeled nucleic acid, as is more fully described below. Thus, the effect of non-specific interactions will be minimized if the label is used to follow the progress of the fractionation. Competitor nucleic acid is also useful in the amplification process. Starting nucleic acid can be prepared with PCR tails to permit amplification, as shown in FIG. 1, step 2. If competitor nucleic acid lacking these PCR tails is added to the mixture prior to amplification, the effect of non-specific interactions will be minimized on PCR amplified heteroduplex nucleic acid because competitor nucleic acid that appears in the heteroduplex mixture will not be amplified.

V. Detection of Nucleotide Pair Mismatches

The detection of heteroduplex nucleic acid according to the invention is accomplished using a binding assay in which one or more mismatch binding protein(s) bind to a nucleotide mismatch to form a nucleic acid/protein complex which is subsequently detected.

For diagnosis of a genetic disease where the mutation that causes the disease is known, the invention provides methods which enable detection of the presence of heteroduplexes between patient and reference nucleic acids. The invention utilizes known methods of nucleic acid hybridization to form duplexes of test and reference strands, and provides inventive methods for the sensitive detection of even a single base pair mismatch in a heteroduplex. Thus, a genetic disease, one example of which is sickle-cell anemia, which involves the substitution of a thymine for an adenine at position 17 of the gene sequence encoding the beta chain of hemoglobin, is easily diagnosed by the mismatch detection methods of the invention, as described below. Other diseases involving genetic mutations which are diagnosable accord-

5,750,335

15

ing to the invention include the following. For example, Tajima et al. (Jour. Biochem. 105:249, 1989) disclose a gGAG→AAG base change which leads to a Glu→Lys amino acid substitution and results in apolipoprotein E (ApoE) deficiency; Hirshhorn et al. (Jour. Clin. Invest. 83:487, 1989) describe a mutation which leads to adenosine deaminase (ADA) deficiency, i.e., a single base change (CCG→CAG) leading to a Pro→Gln amino acid substitution; Jagadees et al. (Jour. Cell. Biol. Suppl. 13B:291, 1989) describe mutations at seven different locations within the FX gene. GAT→AAT resulting in an Asp→Asn substitution at position 58, GTG→ATG resulting in a Val→Met substitution at position 68, GCC→ACC resulting in a Glu→Lys substitution at position 156, TCC→TTC resulting in a Ser→Phe substitution at position 188, GCC→ACC resulting in an Ala→Thr substitution at position 335, and GGG→AGG resulting in a Gly→Arg substitution at position 447, each mutation of which results in a Factor X deficiency; Ginsburg et al. (Proc. Nat. Aca. Sci. 86:3723, 1989) describes two mutations, GTC→GAC and CGG→TGG resulting in Val→Asp and Arg→Trp substitutions at positions 844 and 834, respectively, each of which produces a defective von Willebrand Factor 2a; Matsuura et al. (Jour. Biol. Chem. 264:10148, 1989) describe a mutation which leads to adenylate kinase deficiency (CGG→TGG) leading to an Arg→Trp amino acid substitution; Dilella et al. (Nature 327:333, 1987) describes a mutation within the PAH gene, TCGG→TGG resulting in an Arg→Trp substitution at position 408, which produces the condition known as phenylketonuria; Bock et al. (Biochem. 27:6171, 1988) disclose a CCT→CTT single base change which leads to a Pro→Leu amino acid substitution and results in antithrombin III deficiency; Ohno et al. (Jour. Neurochem. 50:316, 1988) reports on a CGC→CAC mutation resulting in an Arg→His substitution at codon 178 of the HexB gene which produces Tay-Sachs disease; Gibbs et al. (Proc. Nat. Aca. Sci. 86:1919, 1989) discloses mutations at seven different codons of the HPRT gene, TCT→TTA resulting in a Phe→Leu substitution at position 73, TTG→TCG resulting in a Leu→Ser substitution at position 130, GCA→TCA resulting in an Ala→Ser substitution at position 160, CGA→TCA resulting in premature termination of translation at position 169, TTC→GTC resulting in a Phe→Val substitution at position 198, CAT→GAT resulting in a His→Asp substitution at position 203, and TGT→TAT resulting in a Cys→Tyr substitution at position 205, each mutation of which results in HPRT deficiency; and Vulliamy et al. (Proc. Nat. Aca. Sci. 85:5171, 1988) discloses mutations at seven different positions within the G6PDH gene. GAT→AAT resulting in an Asp→Asn substitution at position 58, GTG→ATG resulting in a Val→Met substitution at position 68, AAT→GAT resulting in an Asn→Asp substitution at position 126, GAG→AAG resulting in a Glu→Lys substitution at position 156, TCC→TTC resulting in a Ser→Phe substitution at position 188, GCC→ACC resulting in an Ala→Thr substitution at position 335, and GGG→AGG resulting in a Gly→Arg substitution at position 447, each mutation of which produces a condition known as G6PDH deficiency.

A spot detection assay may be used to detect mismatches, as shown in FIG. 5 and described above. This method allows for the detection of genetic differences between a nucleic acid standard (a reference nucleic acid) and a number of test nucleic acids. Any number of conventional detection methods well-known to those skilled in the art may be used; e.g., direct detection of, e.g., labeled mismatched binding protein, detection of a fluorescent antibody capable of binding the

16

mismatch binding protein, or detection of an antibody conjugated to an enzyme that reacts with a chromogenic substrate.

Also included in the invention are detection methods based on the use of modified nucleic acid and proteins capable of binding the modified nucleic acid. For example, a modified base may occur as part of a mismatched nucleotide pair, and a mismatch binding protein capable of binding to the mismatched pair containing the modified base may be used for detection.

A band shift assay may also be used to detect bound heteroduplex nucleic acid according to the invention, as described above for FIGS. 6 and 7.

Other detection methods useful in the invention are illustrated by way of FIG. 1. Heteroduplexes are formed in step 1 and allowed to bind to mismatch binding protein in step 2. The heteroduplex/mismatch binding protein complexes may then be separated from free nucleic acid by immunoprecipitating the complexes with an antibody specific for the mismatch binding protein in step 3, e.g., using the method of McKay (supra). MutS polyclonal antibodies can be prepared according to conventional antibody preparation procedures using the following procedure.

Purified MutS is electrophoresed on an 8% polyacrylamide gel. After soaking in water 10 min. to remove the SDS, the gel is stained for 10 min in 0.1% coomassie blue in water, and then destained in water. The MutS band is cut out, chopped up into fine pieces with a razor blade. 1 ml of PBS (137 mM NaCl, 2.7 mM KCl, 4.3 mM Na₂HPO₄, 1.4 mM KH₂PO₄, pH 7.3) is added, and the mixture is ground up further by passage through progressively smaller syringes. Rabbits are injected with 500 µg of a mixture of fractions containing the MutS protein. Protein for boosts is prepared in the same way, except that Freund's incomplete adjuvant is used. The rabbits are boosted twice with 100 µg of the MutS fractions, and bled to obtain serum.

The serum is pre-absorbed and used in immunoblotting according to the protocols of Harlow and Lane (1988, "Antibodies, A Laboratory Manual," Cold Spring Harbor Press, CSH, New York), hereby incorporated by reference.

After the immunoprecipitation step, heteroduplex nucleic acid fragments may be optionally isolated for further analysis by performing a phenol extraction to remove the binding protein and anti-binding protein antibody.

Alternatively, other means of detecting bound mismatch binding protein may be used; e.g., the mismatch binding protein itself may be labeled or one strand of the heteroduplex nucleic acid may be labeled and followed into bound nucleic acid, also as described herein. Additional detection techniques are described below as procedures for fractionation; e.g., a mismatch binding protein binding column which binds to mismatch binding protein by virtue of a sequence in the binding protein which is recognized by a moiety on the column.

VI. Affinity Fractionation of Heteroduplexes

The invention also provides for selective enrichment of heteroduplexes within a sample by affinity fractionation of fragments containing mismatches, thereby achieving more sensitive detection of the mismatch(es).

The proportion of heteroduplexes in a sample may be substantially increased using affinity fractionation, as shown schematically in FIG. 3. The mixture containing heteroduplexes is subjected to affinity purification, in which the heteroduplexes are bound to and subsequently eluted from a solid support to which mismatch binding protein is coupled. In FIG. 4, the heteroduplex/mismatch binding protein complexes are selectively retained by a matrix to which any

5,750,335

17

moiety is coupled which can bind the complex, e.g., a binding protein specific- or complex specific-antibody.

In addition to antibody supports in which the antibody binds directly to the mismatch binding protein or the nucleic acid/mismatch binding protein complex, other affinity supports may be used. For example, one can take advantage of the ability of a metal, e.g., nickel, column to bind to histidine residues in a polypeptide using immobilized metal affinity chromatography. A histidine tail, e.g., six histidine residues, may be covalently linked to the amino terminus of the mismatch binding protein, as described by Hochuli et al. (November 1988, *Biotechnology*, p. 1321, hereby incorporated by reference). When the heteroduplex/binding protein complex is applied to a nickel column, the histidine portion of the binding protein will be bound by the column. This procedure is also described in Holuchi et al. (*ibid*).

A histidine-tagged MutS protein may be prepared according to the following procedure. This procedure describes the preparation of a His-MutS protein in which six histidine residues have been added to the amino terminus of the MutS protein. Of course, other His-MutS proteins may be prepared; for example, any desired number of histidine residues may be added to the amino terminus of the MutS protein, provided the resultant His-tagged MutS protein retains its biological activity in binding mismatched nucleic acid and is retainable on a nickel column. If desired, the His-MutS protein can be purified further using a 20 mM–120 mM phosphate gradient on a hydroxyapatite column or on other protein purification known in the art.

Briefly, six histidine residues may be added to the amino terminus of the MutS protein. The MutS gene may be PCR amplified from plasmid DNA containing the gene using PCR primers which anneal to each end of the gene and prime DNA replication. The amplified DNA is then digested with restriction endonucleases to generate a restriction fragment containing MutS-encoding DNA. The MutS-encoding restriction fragment is then cloned into a polylinker site of a plasmid which allows for expression of the inserted DNA by placing the inserted DNA under control of a promoter. Preferably, this promoter is controllable so that MutS gene expression is initiated at a desired point in the cell cycle; e.g., the inducible *E. coli* lac promoter is useful in an *E. coli* host. The muts-encoding clone is then transformed into an appropriate host strain, and a clone is isolated containing MutS-encoding DNA.

The MutS-encoding clone is grown under conditions which do not allow for expression of the MutS gene until a desired optical density of the cell culture is reached. The culture is then induced to produce His-MutS, and the cells grown until they are harvested. The cells are then centrifuged, and the pellets are frozen at –80° C. until ready for use. MutS protein is then purified from the cell pellet as follows. The cell pellet is thawed on ice and resuspended in lysis buffer (20 mM KPO4 pH 7.4, 10 mM betamercaptoethanol, 0.5M KCl, 1 mM PMSF, 200 µg/ml lysozyme). The cells are then disrupted by sonication in an ice water bath. Cell debris is then eliminated by centrifugation at 30,000 rpm for 30 min. The supernatant is filtered through a 0.45 micron filter and applied to a Qiagen (Chatsworth, Calif.) nickel column at a rate of approximately 0.5 ml/min. The column is pre-equilibrated with Buffer D (20 mM KPO4 pH 7.4, 10 mM betamercaptoethanol, 0.5M KCl, 1 mM PMSF). The column is then washed with 75 ml of Buffer D, followed by another 10 ml of Buffer D containing 10 mM imidazole. The protein was eluted with 80 mM imidazole in Buffer D. The recovered protein is then dialyzed against dialysis buffer (20

18

mM KPO4 pH 7.4, 10 mM betamercaptoethanol, 0.5M KCl, 0.1 mM EDTA). The MutS protein containing an amino terminal histidine tail is then ready for use.

Another example of an affinity support is an antibody-bound support in which the antibody recognizes and binds to a flag sequence, i.e., any amino acid sequence (e.g., 10 residues) which the antibody specifically binds to. The flag sequence may be engineered onto the amino terminus of the mismatch binding protein. When the heteroduplex/binding protein complex is applied to the antibody column, the antibody will bind to the flag sequence in the binding protein and thus retain the complex. One embodiment of this technique, known as The Flag Biosystem, is commercially available from International Biotechnologies, Inc. (New Haven, Conn.). Larger flag sequences may be also used, e.g., the maltose binding protein, as described by Ausubel et al., 1992, *supra*.

Alternatively, or in addition to the first fractionation step, the eluted heteroduplex nucleic acid is then recycled one or more times through another affinity binding reaction to refractionate the eluted heteroduplexes and thus remove any remaining non-specifically bound and subsequently eluted homoduplex nucleic acid. The refractionated heteroduplexes are then also subsequently eluted.

Other embodiments of affinity fractionation which are within the scope of the invention include amplification of annealed sample nucleic acid and the addition of competitor nucleic acid, as shown in the figures. For example, the sample nucleic acid may be amplified by PCR after the first affinity binding step, but before the refractionation step. Thus, the bound and eluted heteroduplexes will be amplified and repurified on the affinity support. Elution of the repurified sample nucleic acid should yield relatively pure heteroduplex nucleic acid. In addition, excess competitor nucleic acid (i.e., unlabeled where the sample nucleic acid is labeled, or lacking PCR tails where the sample nucleic acid contains PCR tails) may be added to the sample either prior to or after amplification in order to reduce nonspecific mismatch protein binding to mismatched nucleic acid.

Another fractionation method allows for removal of test-test and/or reference-reference hybrids from a sample prior to analysis. As described generally above and in more detail below, this method provides for differential PCR tailing of duplex fragment ends and thus allows for exponential amplification of test-reference hybrids. Thus, a selective reduction is achieved in the frequency of test-test and reference-reference hybrids within a nucleic acid sample.

This technique, shown schematically in FIG. 10, is useful as an intermediate amplification step and can be performed prior to refractionation to limit affinity purification to test-reference heteroduplexes. A conventional PCR amplification reaction is performed using the experimental conditions disclosed in Lisitsyn, *supra*, such that the test-reference hybrids are the only heteroduplexes to undergo exponential amplification.

In yet another fractionation method useful according to the invention, second-order kinetics of self-association can be used to further enrich sample nucleic acid for fragments that are more prevalent than others (see Wieland et al., 1990, *Proc. Nat. Aca. Sci.* 87:2720, hereby incorporated by reference). After sample nucleic acid is enriched for fragments that contain base pair mismatches, e.g., using MutS affinity fractionation, as described herein, these MutS-binding fragments can be further enriched for the relevant sequence using kinetic-enrichment.

Kinetic-enrichment is based on the following principle. If a population of nucleic acid fragments containing a target

5,750,335

19

subpopulation enriched X times relative to unenriched fragments in the sample is melted and reannealed so that only a small proportion of double-stranded nucleic acid forms, double-stranded target nucleic acid would be present X² times relative to the other sequences present as duplex nucleic acid. To visualize this, consider viral sequences present in excess (ten times more) relative to single-copy β -globin sequences. At early stages of self-reannealing, when 5.0% of the viral sequences are reannealed, only 0.5% of the β -globin sequences will be reannealed. The ratio of the viral sequence to the β -globin sequences in the double-stranded DNA will then be 5% of 10 to 0.5% of 1 (i.e., 100-fold more).

The kinetic-enrichment technique is useful according to the invention as follows. Sample nucleic acid is prepared by combining test and reference nucleic acids under denaturing and reannealing conditions. The sample is then enriched for heteroduplexes thus formed, e.g., by MutS affinity fractionation, as described herein. The MutS-bound heteroduplexes are then telecasted, and the heteroduplex sample kinetically enriched, e.g., is again subjected to denaturation and annealing so that only a small proportion of the sample forms duplexes. Duplexed nucleic acid is then selected as described herein. Because duplex formation will occur at a much higher rate for those fragments that were enriched in the original sample (see Lisityn, supra), the technique serves to further enrich the sample for these fragments.

The fractionation procedure allows for a reduction in the number of homoduplexes in the mixture in the bound fraction; consequently, in the detection or analysis steps, there will be fewer non-specific binding interactions between the mismatch binding protein and homoduplex nucleic acid. The sensitivity of detection and/or quantitation of heteroduplex nucleic acid in a test sample may be further increased by refractionating the eluted sample, or by refractionating the flow-through fractions through repeated affinity steps in which heteroduplexes present either in the eluate or flow-through are selectively retained on the solid support.

After each refractionation binding reaction, bound heteroduplex nucleic acid is eluted and subsequently applied to a fresh or regenerated support. Alternatively, the support may contain a vast excess of binding sites, thus making intermediate elution steps unnecessary.

The solid support useful in the invention may be any one of a wide variety of supports, and may include but is not limited to, synthetic polymer supports, e.g., polystyrene, polypropylene, substituted polystyrene, e.g., aminated or carboxylated polystyrene, polyacrylamides, polyamides, polyvinylchloride, etc.; glass bead, agarose; cellulose, or any material useful in affinity chromatography (see Pharmacia LKB Biotechnology Products Catalog, 1992, Piscataway, N.J., hereby incorporated by reference). The supports may be provided with reactive groups, e.g., carboxyl groups, amino groups, etc., to permit direct linking of the protein to the support. The mismatch binding protein can either be directly crosslinked to the support, or proteins (e.g., antibodies) capable of binding the mismatched binding protein or the nucleic acid/binding protein complex can be coupled to the support.

For example, if the support includes sepharose beads and the mismatch binding protein is coupled to the beads, the binding protein coupled-beads are packed into a column, equilibrated, and the column is subjected to the nucleic acid sample. Under appropriate binding conditions, the protein that is coupled to the beads in the column retains the nucleic acid fragments or the protein/nucleic acid complex which it recognizes. The column is then washed of unbound nucleic

20

acid, and the bound nucleic acid fragments or protein/nucleic acid complexes are eluted according to conventional techniques known in the art, e.g., using a solution containing salt (e.g., KCl), detergent or imidazole, that reduces the binding between the nucleic acid and protein on the support or the protein/nucleic acid complex and the support; e.g. see Scopes, *Protein Purification: Principles and Practice*, 1982, Springer-Verlag, New York, or Ausubel, 1992, *Current Protocols*, supra, both of which are hereby incorporated by reference). Conditions for binding and elution of heteroduplex nucleic acid or heteroduplex/binding protein complexes are typically identical to the conditions described herein for the mismatch binding protein/heteroduplex binding reaction.

The protein may be linked to the support by a variety of techniques including adsorption, covalent coupling, e.g., by activation of the support, or by the use of a suitable coupling agent or the use of reactive groups on the support. Such procedures are generally known in the art and no further details are deemed necessary for a complete understanding of the present invention. Representative examples of suitable coupling agents are dialdehydes, e.g., glutaraldehyde, succinaldehyde, or malonaldehyde; unsaturated aldehyde, e.g., acrolein, methacrolein, or crotonaldehyde; carbodiimides; diisocyanates; dimethyladipimate; and cyanuric chloride. The selection of a suitable coupling agent should be apparent to those of skill in the art from the teachings herein.

Any method that permits the purification of protein/nucleic acid complexes away from free nucleic acid may be used, e.g., at steps 3-5 of FIG. 4. Methods of affinity purification of mismatch binding protein/heteroduplex complexes include immunoprecipitation. See Ausubel, 1992, *Current Protocols*, supra, and Harlow et al., 1988, *Antibodies: A Laboratory Manual*, supra. Alternatively, antibodies to the mismatch binding protein/heteroduplex complex can be attached to any solid support that permits the washing away of free nucleic acid. Alternatively, immobilized metal affinity chromatography may be used to purify histidine-tailed mismatch binding protein that is bound to heteroduplexes.

Additional forms of affinity purification of mismatch binding protein/heteroduplex complexes include the use of nitrocellulose filters that bind protein but not free nucleic acid, or the use of a gel electrophoresis mobility shift nucleic acid-binding assay, both of which are described in Ausubel (1992, supra). For example, the method of the invention shown schematically in FIG. 4 may include a gel mobility shift assay at step 2 of the procedure. Nucleic acid fragments that are bound by mismatch binding protein are identified by their mobility shift. The identified fragments are isolated (steps 4 and 5) by excising them from the gel, and purifying them away from the gel material, as described in Ausubel. VII. Utilization of Heteroduplexes

The inventive methods disclosed herein allow for recovery of nucleic acid fragments containing nucleotide sequence mismatches. Described below are some of the ways in which these recovered fragments may be used. For example, a recovered heteroduplex sample may be used to determine the identity and position of the mismatch by determining the nucleotide sequence of the mismatch region and comparing the sequence with sequence data from reference nucleic acid. Other examples of ways to utilize the isolated heteroduplexes are as follows.

Heteroduplexes may be used to quantitatively determine the fraction of heteroduplex fragments in a mixture and the proportion of mismatch binding protein bound to heteroduplex nucleic acid, and thus may be used to determine the number of fragments containing mismatches within a

5,750,335

21

sample. Labeling of the input test or reference nucleic acids allows for quantitation of label in both the input and output affinity fractionated samples (FIG. 2). Thus, the amount of label present in the output sample may be used to quantitate the number of heteroduplexes relative to the known amount of labeled input sample.

Labeling of the mismatch binding protein (e.g., with ³⁵S-methionine) also allows for detection and optional quantitation of the fraction of heteroduplex fragments in a mixture. For example, as shown in FIG. 5, one method includes immobilizing reference nucleic acid on a solid support, such as a membrane, hybridizing of the immobilized reference nucleic acid to test nucleic acid, exposing the membrane to mismatch binding protein under binding conditions such as those specified herein, and then washing away free mismatch binding protein. Alternatively, test nucleic acid may be immobilized to the support and hybridized to free reference nucleic acid prior to binding.

In addition, a moiety that permits affinity purification of nucleic acids can be used to modify the test or reference nucleic acids for detection; e.g., biotin. After the mixture of modified (e.g., biotin-labeled) nucleic acids is exposed to the mismatch binding protein, the mixture may then be selectively enriched for the nucleic acid/binding protein complexes by affinity purification. During this step, the free nucleic acid and free mismatch binding protein will be washed away. Once the nucleic acid mixture has been separated from free mismatch binding protein, the amount of label present in the bound nucleic acid sample may be used to quantitate the number of heteroduplexes in the mixture. Similarly, the amount of label present in the bound protein may be used to determine the number of mismatches present in the mixture. Alternatively, instead of labeling the mismatch binding protein, other methods for detecting the presence of the mismatch binding proteins can be used for quantitation of mismatches, such as an enzyme-linked immunoassay.

If the goal of the genetic screening method is to identify not only the presence of a nucleotide sequence mismatch between test and reference nucleic acids, but also to determine the nature and location of the mismatch, then the affinity purified heteroduplex nucleic acid can be cloned and sequenced to determine the precise sequences and sequence differences between the test and reference nucleic acids. For example, in the genetic disease hemophilia is caused by many different mutations in a 26,000 base region of nucleic acid in the gene encoding blood clotting factor VIII. Thus, it is not possible to diagnose the disease by identifying a known mutation. However, it is possible to detect the many possible mutations which may be a cause of hemophilia according to the invention. Other genetic diseases, e.g., Huntington's disease, in which neither the nature or location of the mutation which causes the disease is known, may be both diagnosed according to the invention, and also characterized as to the identity (i.e., the nature and/or location) of the underlying mutation.

Differential cloning of genomic nucleic acid can be used with complex nucleic acid samples to eliminate background heteroduplex molecules; i.e., heteroduplexes that are formed when a sample is annealed with itself due to the presence of non-unique sequences. This technique is illustrated schematically in FIG. 8. For example, if nucleic acid A and nucleic acid B are to be compared for nucleotide sequence differences, and both samples are a complex mixture of nucleic acid, when the two samples are combined, and denatured and reannealed, many heteroduplexes will form which are not the A/B heteroduplexes which it is the goal to

22

identify, i.e., which contain one strand from sample A mutated gene X and the other strand from reference B normal gene X. Instead, background heteroduplexes will form which contain strands of non-unique nucleic acid that anneal because they are largely homologous; i.e., A/A or B/B heteroduplexes. This background problems may be reduced using the differential cloning method described above, as follows.

Heteroduplexes from denatured and reannealed A/A nucleic acid and denatured and reannealed B/B nucleic acid may be combined to form the reference nucleic acid. The test nucleic acid (A/B heteroduplexes) will include A DNA and B nucleic acid that is denatured and reannealed together rather than separately. The reference (A/A and B/B) nucleic acid is dephosphorylated to prevent ligation of unwanted heteroduplexes to dephosphorylated vector nucleic acid, and then combined with test nucleic acid (heteroduplexes of A/B nucleic acid) in a ratio of approximately 100 (reference) to 1 (test). The combined mixture is separated by size on an agarose gel and again denatured and reannealed in the gel. In the reannealing process, unique A/B strands are more likely to reanneal than non-unique strands because the latter are more likely to reanneal with excess reference strands. Cloning of the unique A/B test strands will be highly favored due to the inability of dephosphorylated A/A or B/B DNA to ligate to the dephosphorylated vector. The differential cloning technique may be varied as desired using the knowledge of a person of skill in the art.

Alternatively, instead of using differential cloning of genomic DNA, representational difference analysis (RDA) can be used in FIG. 8 (see Lisitsyn et al., supra).

In some circumstances, the goal of the genetic screening may not be to identify the precise mismatch, but to determine the sizes of heteroduplex nucleic acid in an annealed sample identified as containing heteroduplex nucleic acid. The size of a heteroduplex may be determined by agarose gel electrophoresis of affinity purified duplexes. Once the size of heteroduplex fragments are known, size parameters may be used to map the locations of differences in simple nucleic acid samples, such as plasmid DNA or to map the locations or differences in more complex samples via Southern blotting of heteroduplex nucleic acid. Furthermore, where a region of interest is well-defined or where genetic markers are known, other techniques may be used, e.g., Restriction Fragment Length Polymorphism analysis to analyze heteroduplex nucleic acid.

The purified heteroduplex nucleic acid may be used as a probe to screen a genomic library for other sequences of interest. The heteroduplex-containing sample may be further purified by affinity fractionating the heteroduplexes, and/or PCR amplifying the annealed mixture or refractionating the affinity purified heteroduplexes, and cloning the heteroduplex molecules.

In addition, any conventional technique for comparing nucleic acids, e.g., denaturing gradient gel electrophoresis, can be used to further analyze the heteroduplex nucleic acid.

When comparing complex nucleic acid samples, it is important to eliminate background; e.g., false positives, or positive signals generated by reannealing of two different regions within the same test nucleic acid sample that contain some homology and some sequence differences. Background can be eliminated by using controls in which the test nucleic acid or reference nucleic acid is denatured and reannealed with itself. Computer-based assistance can be employed to eliminate these artifacts. For example, a computer can be programmed to examine the digitized images from the gel electrophoresis of reannealed test nucleic acid

5,750,335

23

and/or reannealed reference nucleic acid comparisons, and to remove these artifacts from the digitized gels images resulting from a test/reference heteroduplex comparison. VIII. Detection of Heteroduplex nucleic acid in a Mixture of Excess Competitor nucleic acid

The following experiment demonstrates that a test and a reference nucleic acid sequence may be hybridized and a single base pair differences is detectable. In this example, the nucleotide pair mismatch is known, and the procedure results in detection of mutations in a 16-mer substrate. In addition, 16-mer heteroduplex nucleic acid was fractionated from homoduplex (i.e., fully complementary) nucleic acid. A 16-mer homoduplex control was used to ensure that the method did not fractionate matched nucleic acid to the same degree. Both of the fragments were fractionated in the presence of a large amount of (i.e., excess) competitor nucleic acid to ensure the method could detect mismatches in a background of Nucleic acid.

Nucleic acid samples were prepared as follows. The oligonucleotides DG6R (GAT CCG TCG ACC TGC A), DG4R (CTA GGC AGT TGG ACG T) and DG5 (CTA GGC AGC TGG ACG T) were ordered from Operon Technologies (Alameda, Calif.) and separately resuspended in TE buffer to a concentration of 10 pMol/ul. DG6R was kinased with 5000 Ci/mmol ³²P ATP. Lambda ladder DNA from Bethesda Research Laboratories (Bethesda, Md.) was used as a competitor DNA.

Heteroduplexes were created as follows. 8 pmol of the kinased DG6R and 10 pMol of DG4R in 40 ul of assay buffer were placed in a 70° C. water bath for 10 minutes. The water bath was then switch off and allowed to cool to room temperature to allow the oligonucleotides to anneal. The result of this annealing reaction was called DG-4/6 Het. The same annealing reaction was run between DG-5 and DG-6R, and the result of this reaction was called DG-5/6 Hom. DG-4/6 Het. contains a GT mismatch in place of the GC match present in DG-5/6 Hom.

The MutS protein was over produced, as described by Haber (1988, supra), at 42° C. in MM294 mutS::Tn10 cells that carried the lambda cI857 gene on pSE103 (Ellege et al., 1985, J. Bacteriol. 162:777) and the MutS gene on pGW1825 (Haber 1988, supra), all references of which are hereby incorporated. MutS was purified using the method of Su and Modrich (1986, supra). Dilution buffer for MutS includes 0.02M KPO4 pH 7.4/0.05M KCl/0.1 mM EDTA/1 mM dithiothreitol/0.1 mg/ml bovine serum albumin. The purified and concentrated fraction containing MutS was used in the following experiments. MutS polyclonal antibody was also produced according to the method of Haber (1988, supra). The binding of MutS to heteroduplex nucleic acid was performed in assay buffer, as described above.

Affinity fractionation of heteroduplex nucleic acid was performed as follows. Two binding reactions were incubated on ice for 30 minutes, one containing heteroduplex nucleic acid and a control containing homoduplex DNA. The heteroduplex reaction contained 14.5 pMol of MutS, 200 fmol of DG-4/6 Het, and 2 ug of competitor nucleic acid in a total volume of 20 ul. The control reaction contained 14.5 pMol of MutS, 200 fmol of DG-5/6 Hom, and 2 ug of competitor nucleic acid in a total volume of 20 ul. After 30 minutes on ice, 5 ul of anti-MutS antibody was added to each binding reaction, and the result was incubated on ice for 60 minutes. 10 ul of Staphylococcus aureus cells that had been washed twice in assay buffer were added to both binding reactions (see McKay, 1981, supra) and the result was incubated on ice for an additional 30 minutes. Both reactions were then spun in a microfuge for 3 minutes at 4° C. and the pellet was washed 8 times in assay buffer.

24

The pellet from each binding reaction was counted in a scintillation counter to test for immunoprecipitation of heteroduplex nucleic acid. After normalizing for the total number of counts in each reaction, 53 fold more oligonucleotides precipitated in the heteroduplex reaction than in the homoduplex reaction. Thus, heteroduplexes containing a single base pair mismatch could be detected after affinity fractionation of a mixture containing excess competitor nucleic acid.

IX. Detection of a Mismatched Nucleotide Pair in a 1 KB Fragment

The invention may be used to identify a single base pair change in a 1 KB region of nucleic acid in the presence of an excess of matched nucleic acid competitor.

DNA samples and heteroduplexes were prepared as follows. Single stranded circular DNA from M13mp8 DNA containing a G to A transition mutation in the unique PstI site (see Loechler, 1984, Proc. Nat. Aca. Sci. U.S.A. 80:6271, hereby incorporated by reference) was denatured and annealed in the presence of linear duplex wild-type M13mp8 DNA to create a heteroduplex (see Kramer et al., 1989, J. Bacteriol. 171:5339, hereby incorporated by reference). The heteroduplex thus formed contained a C-A mismatch in the PstI site, which prevented cleavage of the site by PstI. Control homoduplex DNA was created using the sense and antisense strands of wild-type M13mp8 DNA. The 1 KB AvaII-BglIII fragment containing the mismatch was isolated from both the heteroduplex and wild-type homoduplex DNA by gel purification. The resulting homoduplex and heteroduplex fragments were separately phosphatased and end labeled with ³²P ATP. Free ATP was eliminated with spin columns from the labeled heteroduplex and homoduplex 1 KB DNA fragments. Lambda ladder DNA from BRL was used as a competitor.

Affinity fractionation of heteroduplex nucleic acid was performed as follows. Two binding reactions were incubated on ice for 30 minutes, one of which contained the mismatched nucleic acid and a control which contained matched nucleic acid. The heteroduplex-containing reaction consisted of 42 pMol of MutS, 7 fmol of the C-A mismatched 1 KB fragment, and 1 ug of competitor nucleic acid in a total volume of 10 ul. The homoduplex reaction contained the same components, but substituted matched nucleic acid for the mismatched heteroduplex nucleic acid. After 30 minutes on ice, 10 ul of anti-MutS antibody was added to each binding reaction, and the result was incubated on ice for 60 minutes. Then 10 ul of SAC cells that had been washed twice in assay buffer were added to both binding reactions, and the result was incubated on ice for an additional 30 minutes. Both binding reactions were then spun in a microfuge for 3 minutes at 4° C., and the resulting pellet was washed 6 times in assay buffer.

The pellet from each binding reaction was counted in a scintillation counter to test for specific fractionation of heteroduplex nucleic acid. After normalization for the total number of counts in each reaction, 9.6 fold more fragments precipitated in the heteroduplex reaction than in the homoduplex reaction. Thus, a mismatch of a single nucleic acid base pair could be detected in presence of a large amount of competitor nucleic acid.

X. Detection of a Mismatched Nucleotide Pair in a Mixture of Nucleic Acid Fragments

The invention may be used to detect a single nucleotide pair mismatch in a mixture of nucleic acid fragments, as described below.

A mixture of homoduplex and heteroduplex nucleic acid was prepared from purified PstI+ and PstI- M13mp8 DNA.

5,750,335

25

The PstI+ DNA is wild-type M13mp8 DNA, which is cleavable by the restriction enzyme PstI when in double-stranded form, while the PstI- DNA is M13mp8 DNA with a single base C to T mutation in the unique PstI site (the second C in the PstI site is the one that is mutated which prevents cleavage by PstI). 75 ug of both PstI- DNA and PstI+ DNA were separately cleaved with the EcoRI and PvuII restriction enzymes in a total volume of 250 ul each. 200 ul of each reaction were combined, phenol/chloroform extracted, ethanol precipitated, and resuspended in 1x SSC in an eppendorf tube. The tube was boiled in a beaker over a hot water bath for 10 minutes, and then left to cool to 65 degrees for 15 minutes, then moved to a 65 degree water bath, which was switched off and left overnight to cool. The sample was run on a 2% agarose gel, and the 159 bp band was excised. The 159 bp fragments were purified from the gel slice and resuspended in TE buffer. The fragments were then labeled with ³²p dATP in a Klenow fill-in reaction. The unincorporated dATP was eliminated with a spin column. The purified DNA included both heteroduplex and homoduplex nucleic acid.

Mismatch binding protein was bound to the nucleic acid mixture in a total volume of 10 ul consisting of 1 ul of the DNA mixture (19 fMol), 2 ul of the mismatch binding protein MutS (4 ug), and 1 ul of poly dIdC competitor nucleic acid (1 ug). A control reaction was identically prepared except that it did not contain MutS. Binding was performed on ice for 30 minutes. The MutS reaction and the control reaction were electrophoresed on a 6% non-denaturing tris-acrylamide-EDTA (TAE) gel. 2 uL of a 50% sucrose solution was added to each reaction just prior to gel loading.

FIG. 7 shows results from an autoradiogram of the polyacrylamide gel. In lane 1, the control reaction shows a single 159 bp band, while Lane 2 shows both the 159 bp band arising from the homoduplex component of the DNA mixture and a larger molecular weight shift band corresponding to the heteroduplex component of the mixture. Lane 3 shows another control in which the MutS protein was heated prior to the binding reaction. As the results show, heat denatured MutS does not bind to heteroduplex nucleic acid and thus does not result in a band shift in the gel.

XI. Preparation of Histidine-tailed MutS Protein

A variant of the native *Salmonella* MutS protein was created that contained six histidines at its amino terminus to facilitate purification of the His-MutS protein or recovery of the His-MutS protein/heteroduplex nucleic acid complex.

The wild type *Salmonella* MutS gene was PCR amplified from the plasmid pGW1811 using the following primers:

DKG-MUTSST

5' CGG AAT TCG CAT CAT CAT CAT CAT ATG AAT GAG TCA
TTT GAT AAG G (SEQ ID NO. 1)

DKG-MUTS3X

5' CGC GGA TCC TTA CAC CAG ACT TTT CAG CCG (SEQ ID NO. 2)

The amplified nucleic acid fragment was cut with EcoRI and BamHI and cloned into the polylinker site of pUC18, which placed the MutS-encoding DNA under the control of the inducible Lac promoter. The resulting plasmid, called pDKGA1, was used to transform the *E. coli* strain GW3732 (Haber, 1988 supra).

A clone (GW3732 pDKGA1) was isolated which contained the plasmid pDKGA1. Because the Lac expression system permits a moderate level of basal transcription, some His-MutS protein is produced even under conditions which result in repression of the lac promoter. This low level of

26

His-MutS production results in poor growth of the transformed cells, and the selective pressure can result in loss of the plasmid from the transformed cells. Thus, care was taken to ensure that the culture did not grow to high density under selective conditions. The His-MutS protein was prepared and purified as follows.

Two 1 liter cultures of GW3732 cells containing plasmid PDKGA were grown with shaking at 37° C. to an OD₆₀₀ of 0.75. The cultures were then induced to produce His-MutS by adding 1 mM IPTG. The cells were grown for another two hours, and then harvested by centrifugation to a cell pellet, decanting the supernatant, and freezing the pellets at -80° C.

A 500 ml culture pellet was then defrosted on ice and resuspended in lysis buffer (20 mM KPO4 pH 7.4, 10 mM betamercaptoethanol, 0.5M KCl, 1 mM PMSF, 200 ug/ml lysozyme). The cells were sonicated in an ice water bath. Cell debris was eliminated by centrifugation at 30,000 rpm for 30 minutes. The supernatant was filtered through a 0.45 micron filter and applied to a Qiagen nickel column at flow rate of 0.5 ml/minute. The column was pre-equilibrated with Buffer D (20 mM KPO4 pH 7.4, 10 mM betamercaptoethanol, 0.5M KCl, 1 mM PMSF). The column was washed with 75 ml of Buffer D, followed by another 10 ml wash of Buffer D with 10 mM imidazole. The protein was eluted with 80 mM imidazole in Buffer D. The recovered protein was dialyzed against dialysis buffer (20 mM KPO4 pH 7.4, 10 mM betamercaptoethanol, 0.5M KCl, 0.1 mM EDTA). FIG. 9 is a polyacrylamide gel showing results of histidine-tailed MutS purification using an imidazole gradient. The His-MutS protein appears in the purification near the 97 KD marker. Histidine-tailed MutS produced as described above was shown to be biologically active in selective binding to nucleic acid mismatches as follows.

XII. Selective Purification of Heteroduplex Nucleic Acid Using Histidine-tailed MutS Protein

Homoduplex and heteroduplex nucleic acid were prepared as follows. Three oligonucleotides:

SRB-5-G 3' GAC ATC TGA TCC GTC GAC CTG CAG ATG
AAG A 5' (SEQ ID NO. 3)

SRB-3-T 5' CTG TAG ACT AGG CAG TTG GAC GTC TAC
TTC T 3' (SEQ ID NO. 4)

SRB-3-C 5' CTG TAG ACT AGC CAG CTG GAC GTC TAC
TTC T 3' (SEQ ID NO. 5)

were obtained from Operon Technologies (Alameda, Calif.). Each oligonucleotide was resuspended in TE buffer to a concentration of 10 pMol/ul. SRB-3-T was end labeled in a kinase reaction using 5000 Ci/mmol ³²P-ATP.

Heteroduplex nucleic acid was prepared by combining 8 pmol of the kinased SRB-5-G oligonucleotide and 10 pmol of the SRB-3-T oligonucleotide, followed by incubation of the combined oligonucleotides in a 70° C. water bath for 10 minutes. The oligonucleotides were allowed to anneal by switching off the water bath, and allowing it to cool to room temperature. The duplex formed as a result of this annealing reaction was called SRB/HET.

Homoduplex nucleic acid was prepared by combining 8 pmol of the kinased SRB-5-G oligonucleotide and 10 pMol of the SRB-3-C oligonucleotide, and treating the combined oligonucleotides as described above for preparation of heteroduplex SRB/HET. The resultant homoduplex nucleic acid was called SRB/HOM. SRB/HET and SRB/HOM differ in that the heteroduplex nucleic acid contains a GT mismatch in place of a GC match present in the homoduplex nucleic acid.

5,750,335

27

Affinity fractionation of heteroduplex nucleic acid was accomplished by performing a binding reaction between the duplex nucleic acid and the His-MutS mismatch binding protein prepared as described above. Briefly, two binding reactions were performed, one containing heteroduplex nucleic acid and a control containing homoduplex nucleic acid. The heteroduplex reaction contained 200 fmol of SRB/HET and 100 pMol of His-MutS, and binding was performed on ice for 30 minutes in assay buffer (20 mM rKPO₄ pH 7.6, 5 mM MgCl₂, 0.1 mM betamercaptoethanol). The homoduplex binding reaction was performed using 200 fmol of SRB/HOM in place of SRB/HET under the same conditions.

Each reaction was added to 100 µl of Ni-NTA (nickel) resin (Qiagen) in a spin column that had been washed in assay buffer. After addition of the reaction mixtures, each spin column was washed six times with assay buffer containing 1% Triton, and bound DNA was eluted with 1M imidazole, pH 7.0. In the case of the SRB/HET DNA, 27% of the DNA was recovered, while in the case of the SRB/HOM DNA, 2% of the DNA was recovered. The results demonstrate that the His-MutS mismatch protein selectively binds heteroduplex nucleic acid, and that the His-MutS/ heteroduplex nucleic acid complex may be selectively retained via affinity purification on a nickel column.

XIII. Selective Recognition and Purification of Mutations in the ARC Gene using PCR Amplified Nucleic Acid

Heteroduplex and homoduplex nucleic acid were prepared as follows. Plasmids derived from pTA200 containing the wild-type ARC gene and EG36 mutant ARC gene (Vershon et al., *Proteins: Structure, Function and Genetics* 1:302, 1986, hereby incorporated by reference) were isolated and used in separate PCR reactions to amplify a region of the ARC gene. PCR reactions included 100 ng of plasmid DNA, 60 pmol of both of the primers ARC5-1 and ARC3-5, and standard PCR reaction components (i.e., PCR buffer, thermostable DNA polymerase, 2 mM of each oligonucleotide). The primer oligonucleotides have the following sequences:

ARC5-1 CCG CGC GAT GAA AGG AAT GAG (SEQ ID NO. 6)

ARC3-5 GGC TTC AAC TTT ACG CGC CAA (SEQ ID NO. 7).

PCR reaction products from the wild-type and EG36 plasmids were gel purified on a 1.5% TAE (tris-acrylamide EDTA) gel, and the 200 bp band was isolated from both. The gel-purified 200 bp PCR products derived from the wild-type and EG36 plasmids were named ARC-WT and ARC-EG36, respectively.

A mixture of heteroduplex nucleic acid and homoduplex nucleic acid, ARC-WT/EG36 was created as follows. A total of 500 ng of both ARC-WT and ARC-EG36 were combined in a 50 mM KCl solution and boiled for five minutes in a water bath. The sample was then allowed to cool slowly to room temperature, and then gel purified on a 1.5% TAE gel. The resulting DNA contained both homoduplex nucleic acid and heteroduplex nucleic acid with GT and CA mismatches. The DNA was then kinased with ³²P-ATP, and unincorporated ATP was separated using a spin column.

ARC-WT/WT homoduplex nucleic acid was created as follows. A total of 1000 ng of ARC-WT DNA was suspended in a 50 mM KCl solution and boiled for five minutes in a water bath. The sample was then allowed to cool slowly to room temperature, and then gel-purified on a 1.5% TAE gel. The resulting DNA contained homoduplex DNA that had been reannealed. The DNA was then kinased with ³²P-ATP, and unincorporated ATP was separated using a spin column.

28

Affinity purification of heteroduplex DNA was performed as follows. A total of 800 fMol of ARC-WT/EG36 was combined on ice with a final concentration of 0.8 uM His-MutS in assay buffer (20 mM KPO₄ pH 7.4, 5 mM MgCl₂, 0.4 mM A-mercaptoethanol). After incubation for 30 min. on ice, the reaction was added to a spin column of Ni-NTA nickel resin. Before use, the spin column was washed and equilibrated in assay buffer. After the reaction was added to the spin column, the column was washed six times with assay buffer and 1% triton, and eluted with 1M imidazole pH 7.0. An identical affinity purification reaction was performed with ARC-WT/WT. In the case of ARC-WT/EG36, 4% of the DNA was recovered, and in the case of ARC-WT/WT, 2% of the DNA was recovered. The results demonstrate that the His-MutS mismatch protein selectively binds heteroduplex DNA, and that the His-MutS/ heteroduplex DNA complex may be selectively retained via affinity purification.

XIV. Selective Recognition and Purification of Amplified Human Nucleic Acid Containing a Genetic Mutation

A genetic mutation contained within human nucleic acid may be detected as follows. Nucleic acid encoding wild type and mutant human β-globin sequences may be cloned into plasmids as described by Abrams et al., *Genomics* 7:463, 1990, hereby incorporated by reference. The plasmid pEGb0c39, described in Abrams et al., contains a naturally occurring C to T mutation in codon 39 of the β-globin gene; the plasmid pEGwt contains the wild-type sequence. These DNA fragments are amplified by performing large scale plasmid preparation of pEGwt and pEGb0c39. Each amplified DNA is then digested with the restriction enzymes NcoI and BamHI, phenol extracted, and ethanol precipitated.

Heteroduplex nucleic acid is then formed as follows. 25 ug of digested pEGb0c39 and 25 ug of digested pEGwt DNA are combined in a 50 ul volume of 50 mM NaCl (β-Het DNA). The sample is then heated to 99° C. for more than 5 min. and allowed to cool slowly to room temperature. The same reactions is performed using pEGwt DNA to form β-Hom DNA. Each of β-Het and β-Hom are then gel-purified as 438 bp NcoI-BamHI fragments. Purified β-Het fragment is called BO/WT DNA and purified β-Hom is called WT/WT DNA.

Affinity fractionation of heteroduplex nucleic acid is performed as follows. Two binding reactions are incubated on ice for 30 minutes, one containing the BO/WT DNA and a control containing WT/WT DNA. The 14 ul binding reactions contain appropriate amounts of DNA and His-MutS protein. Binding is performed in assay buffer (20 mM Tris-Cl pH 7.6, 5 mM MgCl₂, 0.01 mM EDTA, 0.1 mM DTT). Each binding reaction is added to 100 ul of nickel resin in a spin column that has been washed in assay buffer. The two spin columns are washed six times, and DNA is eluted with 1M imidazole, pH 7.0.

Other Embodiments

Other embodiments are within the following claims.

It is further anticipated that other kinds of mismatches, such as asymmetric methylation, can be detected with proteins that bind to hemi-methylated nucleic acids, such as methyltransferases, e.g., dam.

5,750,335

29

30

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i i i) NUMBER OF SEQUENCES: 7

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 49 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1..49
- (D) OTHER INFORMATION: /note= "DKG-MUTSST PRIMER"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:1:

CGGAATTGCG ATCATCATCA TCATCATATG AATGAATCAT TTGATAAGG

49

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 30 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1..30
- (D) OTHER INFORMATION: /note= "DKG-MUTSXX PRIMER"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:2:

CGCGGATCCT TACACCAGAC TTTTCAGCCG

30

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 31 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1..31
- (D) OTHER INFORMATION: /note= "SRB-S-G"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:3:

AGAAGTAGAC GTCCAGCTGC CTAGTCTACA G

31

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 31 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature

31

5,750,335

32

-continued

(B) LOCATION: 1.31
 (D) OTHER INFORMATION: /note= "SRB-3-T"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:4:

CTGTAGACTA GGCAGTTGGA CGTCTACTTC T

31

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 31 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1.31
- (D) OTHER INFORMATION: /note= "SRB-3-C"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:5:

CTGTAGACTA GGCAGCTGGA CGTCTACTTC T

31

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 28 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1.28
- (D) OTHER INFORMATION: /note= "ARCS-1"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:6:

CCGGCGGATG AAAGGAATGA GCAAAATG

28

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 21 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(i i) MOLECULE TYPE: cDNA

(i x) FEATURE:

- (A) NAME/KEY: misc_feature
- (B) LOCATION: 1.21
- (D) OTHER INFORMATION: /note= "ARCS-5"

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:7:

GGCTTCAACT TTACGCGCCA A

21

I claim:

1. A method of genetic screening for a nucleotide variation, said method comprising:

- (A) providing a test nucleic acid suspected to contain a nucleotide variation and a reference nucleic acid;
- (B) annealing said test and reference nucleic acids under conditions sufficient to produce a mixture comprising a first concentration of heteroduplex and excess homoduplex nucleic acid, wherein said nucleotide variation

comprises one member of a mismatched pair in said heteroduplex, wherein said excess homoduplex nucleic acids are generated by reannealing of a first test or reference nucleic acid strand with a fully complementary second test or reference nucleic acid strand;

- (C) fractionating said heteroduplex from said mixture by affinity purification in which a mismatch repair protein immobilized on a solid support binds said mismatched pair in said heteroduplex;

5,750,335

33

- (D) recovering heteroduplex from said affinity purification to produce a heteroduplex sample which contains a second, higher concentration of said heteroduplex; and
- (E) detecting, as an indication of a genetic variation between said test and reference nucleic acids, the presence of a mismatched nucleotide pair in said sample.
2. A method of enriching a mixture of duplex nucleic acids for heteroduplex nucleic acid, said method comprising:
- (A) providing a mixture of nucleic acids comprising a first concentration of a heteroduplex comprising a test nucleic acid strand and a reference nucleic acid strand, and excess homoduplex nucleic acids, wherein said excess homoduplex nucleic acids are generated by reannealing of a first test or reference nucleic acid strand with a fully complementary second test or reference nucleic acid strand;
- (B) separating said heteroduplex nucleic acid from said mixture by affinity purification in which a mismatch repair protein immobilized on a solid support binds a nucleotide mismatch in said heteroduplex nucleic acid; and
- (C) recovering said heteroduplex nucleic acid from said mismatch repair protein to produce a mixture that contains a second, higher concentration of said heteroduplex.
3. The method of claim 2 wherein step B is conducted by forming a complex between said heteroduplex and said mismatch repair protein and separating said complex from uncomplexed duplex.
4. The method of claim 1 wherein said detecting step comprises detecting one of: said mismatch repair protein bound to said heteroduplex, and said heteroduplex bound to said mismatch repair protein.
5. The method of claim 4 wherein said heteroduplex comprises a detectable moiety and said detecting step comprises detecting said detectable moiety.
6. The method of claim 4 wherein said mismatch repair protein further comprises a detectable moiety and said detecting step comprises detecting said detectable moiety.
7. The method of claim 5 wherein said moiety comprises a label, and said detecting step comprises detecting label bindable by said mismatch repair protein.
8. The method of claim 6 wherein said moiety comprises a label, and said detecting step comprises detecting label bindable to said heteroduplex.
9. The method of claim 4 wherein said detecting step comprises forming an immune complex between one of said bound mismatch repair protein or said bound heteroduplex and an antibody.
10. The method of claim 1 wherein said mismatched nucleotide pair is of unknown identity or location, and further comprising the step of determining the identity or location of said mismatched pair.
11. The method of claim 10 wherein said determining step comprises analyzing the nucleotide sequence of said test or reference nucleic acid of said heteroduplex.
12. The method of claim 1 wherein said steps C and D are repeated prior to performing step E.
13. The method of claim 2 or 3 wherein said steps B and C are repeated prior to performing step E.
14. The method of claim 1 wherein after step (D) but prior to step (E), said method further comprises the additional step of amplifying said heteroduplex comprising said mismatched nucleotide pair.
15. The method of claim 14 wherein said test nucleic acid comprises a first PCR sequence and said reference nucleic acid comprises a second PCR sequence.

34

16. The method of claim 2 or 3 wherein said method further comprises after step (C) the step of amplifying said recovered mixture.
17. The method of claim 16 wherein said test nucleic acid comprises a first PCR sequence and said reference nucleic acid comprises a second PCR sequence.
18. The method of claim 14 wherein said heteroduplex further comprises PCR tails, and said amplifying step comprises performing a polymerase chain reaction.
19. The method of claim 16 wherein said heteroduplex further comprises PCR tails, and said amplifying step comprises performing a polymerase chain reaction.
20. The method of claims 2 or 3 wherein the reference nucleic acid is labeled, said method further comprising the step of, prior to said separating step (B), adding excess unlabeled nucleic acid to said mixture as a competitor, thereby to reduce background.
21. The method of claim 2 or 3 wherein the reference and test nucleic acids comprise PCR tails, and said method further comprises the steps of:
- (i) prior to said separating step, adding excess homoduplex nucleic acid lacking PCR tails; and
- (ii) after said recovering step, amplifying said recovered mixture, thereby to reduce background.
22. The method of claim 2 or 3 wherein said mismatch repair protein comprises a histidine tail.
23. The method of claim 2 or 3 wherein said mismatch repair protein comprises a flag sequence and said solid support comprises an antibody that binds to said flag sequence.
24. A kit for separating a heteroduplex nucleic acid from a mixture of heteroduplex and homoduplex nucleic acids, said kit comprising:
- a solid support on which is immobilized, a mismatch repair protein operative to bind a nucleotide mismatch in said heteroduplex; and
- means for separating said heteroduplex from said mixture.
25. The kit of claim 24 wherein said mismatch repair protein is MutS protein.
26. A kit for separating a heteroduplex nucleic acid from a mixture of heteroduplex and homoduplex nucleic acids, said kit comprising:
- a protein that binds a complex comprising an immobilized mismatch repair protein and a heteroduplex, and
- means for separating said heteroduplex.
27. The kit of claim 24 or 26 further comprising a reference nucleic acid.
28. The kit of claim 24 or 25 wherein said means comprises a buffer suitable for detecting or separating said heteroduplex.
29. The kit of claim 26 wherein said protein capable of binding said mismatch repair protein is immobilized on a solid support.
30. A solid support for preferentially binding heteroduplex nucleic acids, said support comprising:
- a mismatch repair protein immobilized on a solid support and operative to bind a nucleotide mismatch in said heteroduplex.
31. The solid support of claim 30, wherein said mismatch repair protein is MutS protein.
32. The solid support of claim 30 or 31 wherein said solid support comprises an affinity matrix.
33. A method of screening for a nucleotide variation, said method comprising:
- (A) providing a duplex nucleic acid;
- (B) contacting said duplex with a MutS protein immobilized on a solid support and operative to bind a nucleotide mismatch in said duplex; and

5,750,335

35

(C) detecting the binding of said duplex to said immobilized MutS protein as an indication of the presence of said nucleotide variation.

34. A method of screening for a nucleotide variation, said method comprising:

(A) providing a test nucleic acid and a reference nucleic acid;

(B) annealing said test and reference nucleic acids under conditions sufficient to produce a mixture comprising a first concentration of heteroduplex and excess homoduplex nucleic acid, wherein said excess homoduplex nucleic acids are generated by reannealing of a first test or reference nucleic acid strand with a fully complementary second test or reference nucleic acid strand;

(C) fractionating said heteroduplex from said mixture by affinity purification using MutS protein immobilized on a solid support and operative to bind a nucleotide mismatch in said heteroduplex, wherein said MutS protein binds said heteroduplex; and

(D) recovering said bound heteroduplex to produce a heteroduplex sample which contains a second, higher concentration of said heteroduplex, said recovery of heteroduplex being indicative of the presence of said nucleotide variation.

35. A method of enriching a mixture of duplex nucleic acids for heteroduplex nucleic acid, said method comprising:

(A) providing a mixture of heteroduplex nucleic acid and homoduplex nucleic acid;

(B) contacting said mixture with MutS protein immobilized on a solid support and operative to bind a nucleotide mismatch in said heteroduplex, under conditions such that said heteroduplex binds said MutS protein; and

(C) recovering said bound heteroduplex to produce an enriched heteroduplex sample.

36. The method of claim 33, wherein said contacting step is carried out in the presence of excess homoduplex nucleic acid.

37. The method of claim 33, wherein said nucleotide mismatch is at an unknown location or is of unknown identity.

38. The method of claim 33, wherein said duplex is formed by the annealing of a reference nucleic acid and a test nucleic acid.

39. The method of claim 38, wherein said test nucleic acid is suspected of containing a mutation.

40. The method of claim 38, wherein at least one of said test or reference nucleic acids is isolated from an organism.

41. The method of claim 40, wherein said organism is a human.

42. The method of claim 33, wherein at least one nucleic acid strand of said duplex has been amplified prior to duplex formation.

43. The method of claim 33, wherein said duplex comprises a detectable moiety and said detecting step comprises detecting said detectable moiety.

44. The method of claim 33 wherein said detecting step comprises forming an immune complex between one of said MutS protein or said duplex bound in step (B) and an antibody.

45. The method of claim 33 or 34 wherein said nucleotide mismatch is of unknown identity or location, and further comprising the step of determining the identity or location of said nucleotide mismatch.

46. The method of claim 33, wherein after step (B) but prior to step (C), said method further comprises the additional steps of isolating said duplex complexes and amplifying said duplex comprising said nucleotide mismatch.

36

47. The method of claim 34, wherein after step (C) but prior to step (D), said method further comprises the additional steps of isolating said duplex complexes and amplifying said duplex comprising said nucleotide mismatch.

48. The method of claim 46 or 47 wherein said duplex further comprises PCR tails, and said amplifying step comprises performing a polymerase chain reaction.

49. The method of claim 33, 34, or 35, wherein said MutS protein comprises a histidine tail.

50. The method of claim 33, 34, or 35, wherein said MutS protein comprises a flag sequence and said solid support comprises an antibody that binds to said flag sequence.

51. The method of claim 33, 34, or 35, wherein said duplex is further contacted with MutL protein.

52. The method of claim 33, 34, or 35, wherein said duplex is further contacted with MutH protein.

53. A kit for detecting a heteroduplex nucleic acid, said kit comprising:

MutS protein immobilized on a solid support and operative to bind a nucleotide mismatch in said heteroduplex; and

means for detecting said heteroduplex.

54. The kit of claim 53, wherein said MutS protein is labeled.

55. The kit of claim 53, further comprising a first protein that binds said MutS protein.

56. The kit of claim 55, wherein said first protein is labeled.

57. A kit for separating a heteroduplex nucleic acid from a mixture of heteroduplex and homoduplex nucleic acids, said kit comprising:

MutS protein immobilized on a solid support and operative to bind a nucleotide mismatch in said heteroduplex; and

means for separating said heteroduplex.

58. The kit of claim 57, further comprising a protein that binds said MutS protein.

59. The kit of claim 53 or 57, further comprising a reference nucleic acid.

60. The kit of claim 53 or 57, further comprising MutL protein.

61. The kit of claim 53 or 57, further comprising MutH protein.

62. A solid support for preferentially binding heteroduplex nucleic acids, said support comprising:

MutS protein immobilized on said solid support and operative to bind a nucleotide mismatch in said heteroduplex.

63. The solid support of claim 62, wherein said solid support is chosen from a synthetic polymer support, a glass bead, agarose, cellulose, or sepharose.

64. The solid support of claim 62, wherein said solid support further comprises immobilized MutL protein.

65. The solid support of claim 62, wherein said solid support further comprises immobilized MutH protein.

66. The method of claim 1 or 2, wherein said mismatch repair protein is immobilized directly onto said solid support.

67. The method of claim 33, 34 or 35, wherein said MutS protein is immobilized directly onto said solid support.

68. The kit of claim 24 or 26, wherein said mismatch repair protein is immobilized directly onto said solid support.

69. The kit of claim 53 or 57, wherein said MutS protein is immobilized directly onto said solid support.

70. The solid support of claim 62, wherein said MutS protein is immobilized directly onto said solid support.

* * * * *

JS 44 (Rev. 11/04)

CIVIL COVER SHEET

The JS 44 civil cover sheet and the information contained herein neither replace nor supplement the filing and service of pleadings or other papers as required by law, except as provided by local rules of court. This form, approved by the Judicial Conference of the United States in September 1974, is required for the use of the Clerk of Court for the purpose of initiating the civil docket sheet. (SEE INSTRUCTIONS ON THE REVERSE OF THE FORM.)

I. (a) PLAINTIFFS Codon Devices, Inc., Duke University and the Massachusetts Institute of Technology		DEFENDANTS Blue Heron Biotechnology, Inc.	
(b) County of Residence of First Listed Plaintiff _____ (EXCEPT IN U.S. PLAINTIFF CASES)		County of Residence of First Listed Defendant _____ (IN U.S. PLAINTIFF CASES ONLY)	
(c) Attorney's (Firm Name, Address, and Telephone Number) 302-658-9200 Jack B. Blumenfeld (I.D. 1014) Morris, Nichols, Arsht & Tunnell LLP 1201 N. Market Street; Wilm., DE 19801		Attorneys (If Known)	

II. BASIS OF JURISDICTION (Place an "X" in One Box Only)	III. CITIZENSHIP OF PRINCIPAL PARTIES (Place an "X" in One Box for Plaintiff and One Box for Defendant)																
<input type="checkbox"/> 1 U.S. Government Plaintiff <input type="checkbox"/> 2 U.S. Government Defendant <input checked="" type="checkbox"/> 3 Federal Question (U.S. Government Not a Party) <input type="checkbox"/> 4 Diversity (Indicate Citizenship of Parties in Item III)	<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <th style="width: 30%;">Citizen of This State</th> <th style="width: 10%;">PTF</th> <th style="width: 10%;">DEF</th> <th style="width: 50%;">Incorporated or Principal Place of Business In This State</th> </tr> <tr> <td><input type="checkbox"/> 1</td> <td><input type="checkbox"/> 1</td> <td><input type="checkbox"/> 1</td> <td><input type="checkbox"/> 4 <input type="checkbox"/> 4</td> </tr> <tr> <td><input type="checkbox"/> 2</td> <td><input type="checkbox"/> 2</td> <td><input type="checkbox"/> 2</td> <td><input type="checkbox"/> 5 <input type="checkbox"/> 5</td> </tr> <tr> <td><input type="checkbox"/> 3</td> <td><input type="checkbox"/> 3</td> <td><input type="checkbox"/> 3</td> <td><input type="checkbox"/> 6 <input type="checkbox"/> 6</td> </tr> </table>	Citizen of This State	PTF	DEF	Incorporated or Principal Place of Business In This State	<input type="checkbox"/> 1	<input type="checkbox"/> 1	<input type="checkbox"/> 1	<input type="checkbox"/> 4 <input type="checkbox"/> 4	<input type="checkbox"/> 2	<input type="checkbox"/> 2	<input type="checkbox"/> 2	<input type="checkbox"/> 5 <input type="checkbox"/> 5	<input type="checkbox"/> 3	<input type="checkbox"/> 3	<input type="checkbox"/> 3	<input type="checkbox"/> 6 <input type="checkbox"/> 6
Citizen of This State	PTF	DEF	Incorporated or Principal Place of Business In This State														
<input type="checkbox"/> 1	<input type="checkbox"/> 1	<input type="checkbox"/> 1	<input type="checkbox"/> 4 <input type="checkbox"/> 4														
<input type="checkbox"/> 2	<input type="checkbox"/> 2	<input type="checkbox"/> 2	<input type="checkbox"/> 5 <input type="checkbox"/> 5														
<input type="checkbox"/> 3	<input type="checkbox"/> 3	<input type="checkbox"/> 3	<input type="checkbox"/> 6 <input type="checkbox"/> 6														


IV. NATURE OF SUIT (Place an "X" in One Box Only)					
CONTRACT	TORTS	FORFEITURE/PENALTY	BANKRUPTCY	OTHER STATUTES	
<input type="checkbox"/> 110 Insurance <input type="checkbox"/> 120 Marine <input type="checkbox"/> 130 Miller Act <input type="checkbox"/> 140 Negotiable Instrument <input type="checkbox"/> 150 Recovery of Overpayment & Enforcement of Judgment <input type="checkbox"/> 151 Medicare Act <input type="checkbox"/> 152 Recovery of Defaulted Student Loans (Excl. Veterans) <input type="checkbox"/> 153 Recovery of Overpayment of Veteran's Benefits <input type="checkbox"/> 160 Stockholders' Suits <input type="checkbox"/> 190 Other Contract <input type="checkbox"/> 195 Contract Product Liability <input type="checkbox"/> 196 Franchise	PERSONAL INJURY <input type="checkbox"/> 310 Airplane <input type="checkbox"/> 315 Airplane Product Liability <input type="checkbox"/> 320 Assault, Libel & Slander <input type="checkbox"/> 330 Federal Employers' Liability <input type="checkbox"/> 340 Marine <input type="checkbox"/> 345 Marine Product Liability <input type="checkbox"/> 350 Motor Vehicle <input type="checkbox"/> 355 Motor Vehicle Product Liability <input type="checkbox"/> 360 Other Personal Injury CIVIL RIGHTS <input type="checkbox"/> 441 Voting <input type="checkbox"/> 442 Employment <input type="checkbox"/> 443 Housing/Accommodations <input type="checkbox"/> 444 Welfare <input type="checkbox"/> 445 Amer. w/Disabilities - Employment <input type="checkbox"/> 446 Amer. w/Disabilities - Other <input type="checkbox"/> 440 Other Civil Rights	PERSONAL INJURY <input type="checkbox"/> 362 Personal Injury - Med. Malpractice <input type="checkbox"/> 365 Personal Injury - Product Liability <input type="checkbox"/> 368 Asbestos Personal Injury Product Liability PERSONAL PROPERTY <input type="checkbox"/> 370 Other Fraud <input type="checkbox"/> 371 Truth in Lending <input type="checkbox"/> 380 Other Personal Property Damage <input type="checkbox"/> 385 Property Damage Product Liability	<input type="checkbox"/> 610 Agriculture <input type="checkbox"/> 620 Other Food & Drug <input type="checkbox"/> 625 Drug Related Seizure of Property 21 USC 881 <input type="checkbox"/> 630 Liquor Laws <input type="checkbox"/> 640 R.R. & Truck <input type="checkbox"/> 650 Airline Regs. <input type="checkbox"/> 660 Occupational Safety/Health <input type="checkbox"/> 690 Other LABOR <input type="checkbox"/> 710 Fair Labor Standards Act <input type="checkbox"/> 720 Labor/Mgmt. Relations <input type="checkbox"/> 730 Labor/Mgmt. Reporting & Disclosure Act <input type="checkbox"/> 740 Railway Labor Act <input type="checkbox"/> 790 Other Labor Litigation <input type="checkbox"/> 791 Empl. Ret. Inc. Security Act	<input type="checkbox"/> 422 Appeal 28 USC 158 <input type="checkbox"/> 423 Withdrawal 28 USC 157 PROPERTY RIGHTS <input type="checkbox"/> 820 Copyrights <input checked="" type="checkbox"/> 830 Patent <input type="checkbox"/> 840 Trademark SOCIAL SECURITY <input type="checkbox"/> 861 HIA (1395ff) <input type="checkbox"/> 862 Black Lung (923) <input type="checkbox"/> 863 DIWC/DIWW (405(g)) <input type="checkbox"/> 864 SSID Title XVI <input type="checkbox"/> 865 RSI (405(g)) FEDERAL TAX SUITS <input type="checkbox"/> 870 Taxes (U.S. Plaintiff or Defendant) <input type="checkbox"/> 871 IRS—Third Party 26 USC 7609	<input type="checkbox"/> 400 State Reapportionment <input type="checkbox"/> 410 Antitrust <input type="checkbox"/> 430 Banks and Banking <input type="checkbox"/> 450 Commerce <input type="checkbox"/> 460 Deportation <input type="checkbox"/> 470 Racketeer Influenced and Corrupt Organizations <input type="checkbox"/> 480 Consumer Credit <input type="checkbox"/> 490 Cable/Sat TV <input type="checkbox"/> 810 Selective Service <input type="checkbox"/> 850 Securities/Commodities/Exchange <input type="checkbox"/> 875 Customer Challenge 12 USC 3410 <input type="checkbox"/> 890 Other Statutory Actions <input type="checkbox"/> 891 Agricultural Acts <input type="checkbox"/> 892 Economic Stabilization Act <input type="checkbox"/> 893 Environmental Matters <input type="checkbox"/> 894 Energy Allocation Act <input type="checkbox"/> 895 Freedom of Information Act <input type="checkbox"/> 900 Appeal of Fee Determination Under Equal Access to Justice <input type="checkbox"/> 950 Constitutionality of State Statutes

V. ORIGIN (Place an "X" in One Box Only)						
<input checked="" type="checkbox"/> 1 Original Proceeding	<input type="checkbox"/> 2 Removed from State Court	<input type="checkbox"/> 3 Remanded from Appellate Court	<input type="checkbox"/> 4 Reinstated or Reopened	<input type="checkbox"/> 5 Transferred from another district (specify)	<input type="checkbox"/> 6 Multidistrict Litigation	<input type="checkbox"/> 7 Appeal to District Judge from Magistrate Judgment

VI. CAUSE OF ACTION	Cite the U.S. Civil Statute under which you are filing (Do not cite jurisdictional statutes unless diversity):
	35 U.S.C. § 271
	Brief description of cause: Patent infringement

VII. REQUESTED IN COMPLAINT:	CHECK IF THIS IS A CLASS ACTION UNDER F.R.C.P. 23 <input type="checkbox"/>	DEMAND \$	CHECK YES only if demanded in complaint: JURY DEMAND: <input checked="" type="checkbox"/> Yes <input type="checkbox"/> No
------------------------------	--	-----------	--

VIII. RELATED CASE(S) IF ANY	(See instructions): JUDGE	DOCKET NUMBER
------------------------------	---------------------------	---------------

DATE March 14, 2007	SIGNATURE OF ATTORNEY OF RECORD 			
FOR OFFICE USE ONLY				
RECEIPT #	AMOUNT	APPLYING IFP	JUDGE	MAG. JUDGE

AO FORM 85 RECEIPT (REV. 9/04)

United States District Court for the District of Delaware

Civil Action No. 07 - 148

ACKNOWLEDGMENT
OF RECEIPT FOR AO FORM 85

NOTICE OF AVAILABILITY OF A
UNITED STATES MAGISTRATE JUDGE
TO EXERCISE JURISDICTION

I HEREBY ACKNOWLEDGE RECEIPT OF 1 COPIES OF AO FORM 85.

MAR 14 2007

(Date forms issued)



(Signature of Party or their Representative)

Aaron Johnston

(Printed name of Party or their Representative)

Note: Completed receipt will be filed in the Civil Action